Optimum Experimental Designs for Models with a Skewed Error Distribution: with an Application to Stochastic Frontier Models

Mery Helena Thompson

A dissertation submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

Department of Statistics

January 2008

Abstract

Keywords: optimum experimental design, linear design, nonlinear design, parameter dependent, D_A-optimality, C-optimality, approximated information matrices, singular information matrices, multiplicative algorithm, skewed or asymmetrically distributed error, composed error, stochastic frontier model, economic efficiency.

In this thesis, optimum experimental designs for a statistical model possessing a skewed error distribution are considered, with particular interest in investigating possible parameter dependence of the optimum designs. The skewness in the distribution of the error arises from its assumed structure. The error consists of two components (i) *random error*, say V, which is symmetrically distributed with zero expectation, and (ii) some type of *systematic error*, say U, which is asymmetrically distributed with nonzero expectation. Error of this type is sometimes called 'composed' error. A stochastic frontier model is an example of a model that possesses such an error structure. The systematic error, U, in a stochastic frontier model represents the economic efficiency of an organisation.

Three methods for approximating information matrices are presented. An approximation is required since the information matrix contains complicated expressions, which are difficult to evaluate. However, only one method, 'Method 1', is recommended because it guarantees nonnegative definiteness of the information matrix. It is suggested that the optimum design is likely to be sensitive

to the approximation.

For models that are linearly dependent on the $\boldsymbol{\beta}$ parameters, the information matrix is independent of the $\boldsymbol{\beta}$ parameters but depends on the variance parameters of the random and systematic error components. Consequently, the optimum design is independent of $\boldsymbol{\beta}$ but *may* depend on the variance parameters. Thus, designs for linear models with skewed error may be parameter dependent. For nonlinear models, the optimum design may be parameter dependent in respect of both the variance and other parameters, which we will denote by $\boldsymbol{\beta}$.

The information matrix is rank deficient. As a result, only subsets or linear combinations of the parameters are estimable. The rank of the partitioned information matrix is such that designs are only admissible for optimal estimation of the $\boldsymbol{\beta}$ parameters, *excluding* any constant term β_0, plus one linear combination of the variance parameters *and* β_0. The linear model is shown to be equivalent to the usual linear regression model, but with a shifted intercept, say β_0^{**}. This suggests that the admissible designs should be optimal for estimation of the $\boldsymbol{\beta}$ parameters, *excluding* β_0, plus the shifted intercept β_0^{**}.

The shifted intercept β_0^{**} can be viewed as a transformation of the intercept β_0 in the usual linear regression model. Since D_A-optimum designs are invariant to linear transformations of the parameters, the D_A-optimum design for the asymmetrically distributed linear model is just the linear, parameter independent, D_A-optimum design for the usual linear regression model with nonzero intercept. C-optimum designs are not invariant to linear transformations. However, if interest is in optimally estimating the $\boldsymbol{\beta}$ parameters, *excluding* β_0, the linear transformation of β_0 to β_0^{**} is no longer a consideration and the C-optimum design is just the linear, parameter independent, C-optimum design for the usual linear regression model with nonzero intercept. If interest is in estimating the $\boldsymbol{\beta}$ parameters, *and* the shifted intercept β_0^{**}, the C-optimum design will depend on (i) the design region; (ii) the distributional assumption on U; (iii) the matrix

used to define admissible linear combinations of parameters; (iv) the variance parameters of U and V; (v) the method used to approximate the information matrix.

Some numerical examples of designs for a cross-sectional log-linear Cobb-Douglas stochastic production frontier model are presented to demonstrate the nonlinearity of designs for models with a skewed error distribution. Torsney's (1977) multiplicative algorithm was implemented in finding the optimum designs.

Acknowledgements

I would like to acknowledge the Department of Statistics at the University of Glasgow for their generous financial support and for providing me with the opportunity to teach at tertiary level through a Glasgow University Teaching Assistant Scholarship.

Thanks to my parents who provided me with the support and opportunity to commence my tertiary studies. Also to my siblings for giving their youngest sister a relatively sane family life and who keep me in check with the rest of society. To my husband for his love, patience, and encouragement, and also for the many mathematical conversations, emails and text messages, especially on algebra. I would also like to thank my friends and volleyball team mates, especially the postgraduates in the Department of Statistics at the University of Glasgow. Thanks to all the academics and postgraduates from various universities in various countries who provided advice, suggestions and stimulating conversation.

Finally, to my supervisor, mentor and 'statistical father', Dr. Ben Torsney, who not only helped me to develop my research skills but also nurtured my academic career ... and who is always good for a laugh and a pint!

M. Helen Thompson

To my family,

my parents, Alan and Kris, my siblings, Fredy and Rebika,

and my husband,

Dave.

Contents

Appendices

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Models with Skewed Error

Throughout this dissertation, the distinction is made between a random variable, which is written in upper case, and its realisation, written in the corresponding lower case letter. For example, the response Y is a random variable until it takes its realised value y, which is the observed response. The usual statistical model, with random error denoted by V, is written

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) + V.$$

Typically the random error is symmetrically distributed with $\mathbb{E}[V] = 0$, giving expected response

$$\mathbb{E}[Y] = f(\boldsymbol{x}, \boldsymbol{\beta}).$$

An additional assumption under maximum likelihood estimation is that random error is normally distributed as $N(0, \sigma^2)$. Much research has been carried out on optimum designs for both linear and nonlinear forms of this statistical model.

In this thesis we consider optimum designs for a statistical model with *skewed* error, say E, which has nonzero expectation and is not normally distributed.

Clearly, the skewness implies that the error E cannot represent statistical noise alone, if noise is assumed to be symmetrically distributed. The error term E is composed of two components, U and V. A general structure of the 'composed' error is

$$E = c_u U + c_v V, \qquad \mathbb{E}[U] \neq 0,\ \mathbb{E}[V] = 0,\ \{c_u, c_v\} \in \mathbb{R}.$$

That is, it is a linear combination of the two components U and V. A simpler and notationally less cumbersome specification would be to consider an error structure $E = V \pm U$. Any other linear combination of the error components could then be treated as a transformation of variables. The nominal generality of the error specification, in terms of the linear combination given in the equation above, is implemented as a convenient device for the reader. It provides an alternative method for deriving the required density functions by simply substituting in values of c_u and c_v rather than applying an appropriate transformation of variables. Here V is a symmetrically distributed random error attributable to statistical noise, hence it has zero expectation. The component U is some type of systematic error, free from statistical noise, that is asymmetrically distributed and has nonzero expectation. The skewness in the error term U causes skewness in the overall composed error E, in the same direction as U. The statistical model with this asymmetrically distributed error structure is written

$$\begin{aligned} Y &= f(\boldsymbol{x}, \boldsymbol{\beta}) + E \\ &= f(\boldsymbol{x}, \boldsymbol{\beta}) + c_u U + c_v V, \end{aligned}$$

with expected response given by

$$\begin{aligned} \mathbb{E}[Y] &= f(\boldsymbol{x}, \boldsymbol{\beta}) + \mathbb{E}[E] \\ &= f(\boldsymbol{x}, \boldsymbol{\beta}) + c_u \mathbb{E}[U]. \end{aligned}$$

Interest is in exploring optimum designs for this model, with particular interest in investigating possible nonlinearity of the optimum designs.

A special case of this model is an econometric model called a 'stochastic frontier model', used to measure the economic efficiency of organisations. The asymmetrically distributed systematic error, U, in a frontier model represents the efficiency of organisations. In recent years there has been a renewed demand for efficiency analysis. The United Kingdom Government has emphasised the importance of measuring output, productivity and efficiency of public sector organisations. The 2004 Spending Review (HM Treasury 2004) details how the Government has responded to the Gershon Review (Gershon 2004) of public sector efficiency and outlines efficiency targets to be achieved between 2005-2008. The Government's commitment to maximising efficiency within the public sector is a key element in this agenda. Recommendations from the Atkinson Review (Atkinson 2005) of the measurement of government output and productivity have also been incorporated in the 2004 Spending Review and in July 2005, the Office for National Statistics (ONS) launched the United Kingdom Centre for the Measurement of Government Activity (UKCeMGA) to take forward the Atkinson agenda. In this thesis, optimum designs for frontier models are investigated for the case where the error structure is $E = V - U$. This corresponds to a 'single-output cross-sectional log-linear Cobb-Douglas stochastic production frontier model' used to measure output-oriented technical efficiency of organisations.

1.2 Outline of this Dissertation

Optimum designs for a model with an asymmetrically distributed skewed composed error have not been developed in the statistics or econometrics literature. We present some theoretical and numerical results of such designs within this dissertation.

In Chapter 2, the information matrix used for finding optimum designs is derived under four possible specifications of the error distribution. The four

distributions considered for the systematic error, U, are nonnegative half normal, exponential, nonnegative truncated normal, and gamma. These are the most common distributions implemented in the econometrics literature. In all cases the random error, V, is distributed as $N(0, \sigma_v^2)$.

Chapter 3 investigates three possible methods for approximating the information matrix. Approximation methods are considered because the information matrix involves expressions that are difficult to evaluate. Only one of the approximation methods, 'Method 1', is recommended to guarantee nonnegative definiteness of the information matrix.

An overview of the measurement of economic efficiency is presented in Chapter 4. This chapter includes varying classifications of efficiency, descriptions of parametric and nonparametric methods for analysing efficiency models, derivation of formulae for calculating efficiency measures, derivation of information matrices for cross-sectional data, and a discussion on extensions to cross-sectional models.

Chapter 5 provides the theoretical background to optimum experimental designs. A distinction is made between linear and nonlinear design problems, and parameter dependence of designs for the asymmetrically distributed model is established. Continuous and exact design measures are defined, with the focus in this dissertation on continuous optimum designs. Conditions of optimality are given, prefaced with definitions of the Gâteaux and Frèchet directional derivatives used to determine optimality. Several optimality criteria and their derivatives are also given. Since the information matrix for the model of interest is rank deficient, designs with singular information matrices are considered. The issue of invertibility of the information matrix is also dealt with in providing alternative choices of generalised inverses with some results given for partitioned matrices.

The chapter concludes with a description of an algorithm used for finding optimum designs.

Although Chapter 6 gives theoretical results for optimum designs for stochastic production frontier models, the theory more generally pertains to an asymmetrically distributed linear model. Effects of linear transformations of the parameters on the optimum design are reviewed before establishing the equivalence of the the usual linear regression model and the linear model with skewed error through a transformation of the parameter space. The structure and rank of the partitioned singular information matrix are explored in determining admissible designs and some theoretical and numerical results for D_A- and C-optimality are presented.

A summary of the conclusions and suggestions for future research are given in Chapter 7.

Chapter 2

A General Statistical Model with Two Error Terms

2.1 Equation for the Statistical Model

Consider the standard statistical model

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) + V,$$

where the observed response Y is a real-valued random variable. The true response is $f(\boldsymbol{x}, \boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is a vector of p unknown parameters and $\boldsymbol{x}$ is a vector of m explanatory variables. The true response $f(\boldsymbol{x}, \boldsymbol{\beta})$ is subject to random error V, giving the observed response Y. It is usual to assume that V is normally distributed, i.e. the distribution of V is symmetric.

Suppose the error in the model is not symmetric. If random error is assumed to be symmetrically distributed then some other process, apart from random error, must also be occurring. The error due to this other process must be asymmetrically distributed for the overall error in the model to be asymmetric. So the overall (asymmetric) error in the model can be modelled as a linear combination

of a symmetric random error term V and an asymmetric error term U

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) + c_u U + c_v V, \qquad \{c_u, c_v\} \in \mathbb{R}. \tag{2.1}$$

This type of model can be found in the econometrics literature typically with $c_u = \pm 1$ and $c_v = 1$. Here we consider the more general case where c_u and c_v can take any real value. Because error terms U and V are unobserved quantities, the contribution from each to the overall error is unknown. Although the two error terms cannot be observed separately, we shall see later that their moments can be calculated separately, conditional on the overall error.

If we let random variable E be the combined error such that $E = c_u U + c_v V$ then model (2.1) becomes

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) + E. \tag{2.2}$$

Different asymmetric distributions for U are considered in the following sections. Section 2.2 considers the case where U has a nonnegative half normal distribution. In Section 2.3, random variable U has an exponential distribution. A generalisation of the half normal distribution is considered in Section 2.4 where U is distributed with a nonnegative truncated normal distribution. Finally, Section 2.5 explores a generalisation of the exponential distribution where U is gamma distributed. These are the distributions that dominate the econometrics literature when the overall error in a stochastic model is a combined asymmetric term such as $E = c_u U + c_v V$ above. Symmetric random error V will be normally distributed with mean zero and constant variance σ_v^2.

2.2 Normal-Half Normal Model

Assume that random variables U and V are distributed as follows

(i) $U \sim N^+(0, \sigma_u^2)$ i.i.d., i.e. nonnegative half normal

(ii) $V \sim N(0, \sigma_v^2)$ i.i.d.

(iii) U and V are distributed independently of each other.

The nonnegative half normal distribution considered here is the normal distribution truncated from below at $\mu = 0$. It is a special case of the nonnegative truncated normal distribution which is discussed in Section 2.4. Appendix C.5 provides further details on truncated normal distributions.

The probability density functions of U and V are

$$f_U(u; \sigma_u) = \frac{2}{\sqrt{2\pi}\sigma_u} \exp\left\{-\frac{u^2}{2\sigma_u^2}\right\}, \qquad u \geq 0,\ \sigma_u > 0, \tag{2.3}$$

$$f_V(v; \sigma_v) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{v^2}{2\sigma_v^2}\right\}, \qquad -\infty < v < \infty,\ \sigma_v > 0, \tag{2.4}$$

with U having mean and variance

$$\begin{aligned} \mathbb{E}[U] &= \sqrt{\frac{2}{\pi}}\sigma_u, \\ Var(U) &= \frac{\pi - 2}{\pi}\sigma_u^2, \end{aligned}$$

and V having mean and variance

$$\begin{aligned} \mathbb{E}[V] &= 0, \\ Var(V) &= \sigma_v^2. \end{aligned} \tag{2.5}$$

Three different normal distributions are plotted in Figure 2.1, and Figure 2.2 depicts three different half-normal distributions.

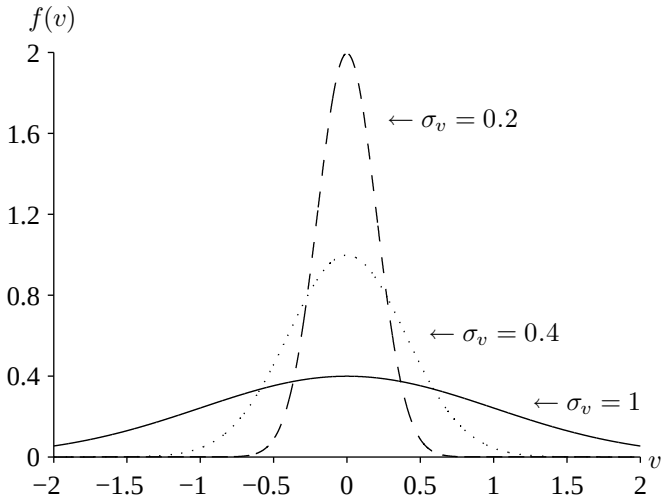

Figure 2.1: Normal distributions.

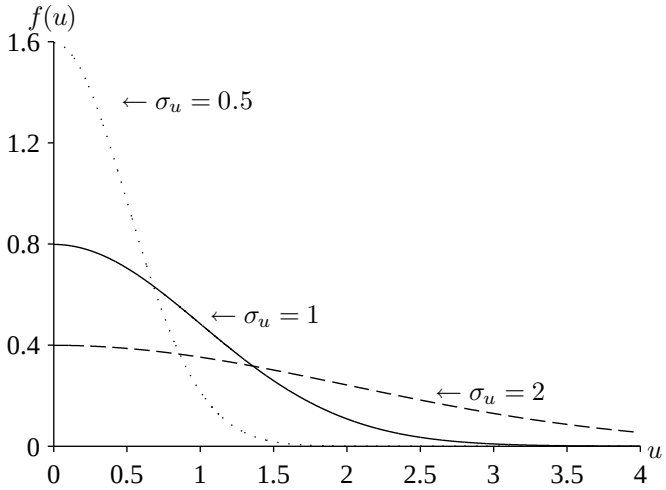

Figure 2.2: Half normal distributions.

The joint probability density function of U and V is

$$\begin{aligned} f_{U,V}(u,v) &= f_U(u) \cdot f_V(v) \\ &= \frac{1}{\pi\sigma_u\sigma_v} \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\}. \end{aligned} \tag{2.6}$$

For random variable E, where $E = c_u U + c_v V$ and $\{c_u, c_v\} \in \mathbb{R}$, the joint density function of U and E can be derived using equation (C.1) in Appendix C and is given by

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= \frac{1}{|c_v|} f_{U,V}\left(u, \frac{\varepsilon - c_u u}{c_v}\right) \\ &= \frac{1}{|c_v|\pi\sigma_u\sigma_v} \exp\left\{-\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon - c_u u)^2}{2c_v^2\sigma_v^2}\right\} \\ &= \frac{1}{|c_v|\pi\sigma_u\sigma_v} \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_u^2} + \frac{c_u^2}{c_v^2\sigma_v^2}\right)u^2 - 2\frac{c_u\varepsilon}{c_v^2\sigma_v^2}u + \frac{\varepsilon^2}{c_v^2\sigma_v^2}\right]\right\}. \end{aligned} \tag{2.7}$$

If we let $K = \dfrac{1}{|c_v|\pi\sigma_u\sigma_v}$, $A = \dfrac{1}{\sigma_u^2} + \dfrac{c_u^2}{c_v^2\sigma_v^2}$, $B = \dfrac{c_u\varepsilon}{c_v^2\sigma_v^2}$ and $C = \dfrac{\varepsilon^2}{c_v^2\sigma_v^2}$ then the joint density function of U and E becomes

$$f_{U,E}(u,\varepsilon) = K \exp\left\{-\frac{1}{2}\left[Au^2 - 2Bu + C\right]\right\}.$$

When the joint density of U and E is of this form, the marginal density of E is given by equation (C.6) in Appendix C as

$$f_E(\varepsilon) = K\sqrt{\frac{2\pi}{A}} \exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \Phi\left(\frac{B}{\sqrt{A}}\right).$$

If we let $\sigma_G^2 = c_u^2\sigma_u^2 + c_v^2\sigma_v^2$ and $\lambda = \sigma_u/\sigma_v$ then

$$A = \frac{\sigma_G^2}{c_v^2\sigma_u^2\sigma_v^2},$$

$$K\sqrt{\frac{2\pi}{A}} = \frac{1}{\sigma_G}\sqrt{\frac{2}{\pi}},$$

$$C - \frac{B^2}{A} = \frac{\varepsilon^2}{\sigma_G^2},$$

$$\frac{B}{\sqrt{A}} = \frac{c_u \lambda \varepsilon}{|c_v|\sigma_G}.$$

The marginal density of E is then given by

$$\begin{aligned} f_E(\varepsilon) &= \frac{1}{\sigma_G}\sqrt{\frac{2}{\pi}} \exp\left\{-\frac{\varepsilon^2}{2\sigma_G^2}\right\} \Phi\left(\frac{c_u \lambda \varepsilon}{|c_v|\sigma_G}\right) \\ &= \frac{2}{\sigma_G}\phi\left(\frac{\varepsilon}{\sigma_G}\right) \Phi\left(\frac{c_u \lambda \varepsilon}{|c_v|\sigma_G}\right), \end{aligned} \tag{2.8}$$

with mean and variance that can be derived using equations (C.7) and (C.8) in Appendix C and which are given by

$$\mathbb{E}[E] = \tilde{c_u}\sigma_u, \tag{2.9}$$

$$Var(E) = \tilde{\tilde{c_u}}^2 \sigma_u^2 + c_v^2 \sigma_v^2, \tag{2.10}$$

where $\tilde{c_u} = c_u\sqrt{\frac{2}{\pi}}$ and $\tilde{\tilde{c_u}}^2 = c_u^2\frac{\pi - 2}{\pi} = c_u^2 - \tilde{c_u}^2$.

The conditional density of U given E can be calculated using equation (C.12) and is given by

$$f_{U|E}(u|\varepsilon) = \frac{\sqrt{A}\phi\left(\frac{u - B/A}{1/\sqrt{A}}\right)}{\Phi\left(\frac{B}{\sqrt{A}}\right)}.$$

The expected value and mode of U given E can be calculated using equations (C.13) and (C.14) respectively and are given by

$$\mathbb{E}[U|E] = \frac{B}{A} + \frac{\frac{1}{\sqrt{A}}\phi\left(-\frac{B/A}{1/\sqrt{A}}\right)}{1 - \Phi\left(-\frac{B/A}{1/\sqrt{A}}\right)},$$

$$M(U|E) = \frac{B}{A}.$$

2.2.1 Log-likelihood function

Under model (2.2), the log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \sigma_G)$ for a sample of N independent observations can be obtained using equation (2.8) and is given by

$$\begin{aligned} \ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{i=1}^{N} \ln f_{Y_i}(y_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \ln f_{E_i}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}); \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \left\{ \ln\left(\frac{2}{\sigma_G}\right) + \ln \phi\left(\frac{y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\sigma_G}\right) + \ln \Phi(-a_i) \right\}, \end{aligned} \tag{2.11}$$

where

$$\begin{aligned} a_i &= -\frac{c_u \lambda \varepsilon_i}{|c_v| \sigma_G} \\ &= -\frac{c_u \lambda [y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})]}{|c_v| \sigma_G}. \end{aligned}$$

Appendix D.1 provides further information on likelihood functions. To reduce notational clutter, observation subscripts will henceforth be omitted.

The expected value and variance of a, which will be used in later chapters, are

$$\mathbb{E}[a] = -\frac{c_u \lambda}{|c_v| \sigma_G} \mathbb{E}[E],$$

$$Var(a) = \left(\frac{c_u \lambda}{|c_v| \sigma_G}\right)^2 Var(E),$$

where $\mathbb{E}[E]$ and $Var(E)$ are given in equations (2.9) and (2.10) respectively. The derivative of a with respect to the parameter vector $\boldsymbol{\beta}$, which will also be used in later chapters, is

$$\frac{\partial a}{\partial \boldsymbol{\beta}} = \left(\frac{c_u \lambda}{|c_v| \sigma_G}\right) \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

In the following equations, some of the derivatives involve the term $y - f(\boldsymbol{x}, \boldsymbol{\beta})$, which can be reparameterised as a function of a. Expressing the derivatives as functions of a simplifies the approximations that will be applied in Chapter 3. The first-order derivatives of $\ln f_Y(y; \boldsymbol{\theta})$ are

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= -\left\{-\frac{y - f(\boldsymbol{x}, \boldsymbol{\beta})}{\sigma_G^2} + \frac{c_u \lambda}{|c_v| \sigma_G} h(a)\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -\left\{\frac{|c_v|}{c_u \lambda \sigma_G} a + \frac{c_u \lambda}{|c_v| \sigma_G} h(a)\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \\
\frac{\partial \ln f_Y}{\partial \lambda} &= \frac{c_u [y - f(\boldsymbol{x}, \boldsymbol{\beta})]}{|c_v| \sigma_G} h(a) \\
&= -\frac{1}{\lambda} a h(a), \\
\frac{\partial \ln f_Y}{\partial \sigma_G^2} &= -\frac{1}{2\sigma_G^2} + \frac{[y - f(\boldsymbol{x}, \boldsymbol{\beta})]^2}{2\sigma_G^4} - \frac{c_u \lambda [y - f(\boldsymbol{x}, \boldsymbol{\beta})]}{2|c_v| \sigma_G^3} h(a) \\
&= -\frac{1}{2\sigma_G^2} + \frac{1}{2}\left(\frac{|c_v|}{c_u \lambda \sigma_G}\right)^2 a^2 + \frac{1}{2\sigma_G^2} a h(a),
\end{aligned}$$

where $h(.)$ is the normal hazard function. Appendix C.6 provides further details on hazard functions and their derivatives. The corresponding second-order derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \left\{-\frac{1}{\sigma_G^2} + \left(\frac{c_u \lambda}{|c_v| \sigma_G}\right)^2 h(a)[a - h(a)]\right\} \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \\
&\quad - \left\{\frac{|c_v|}{c_u \lambda \sigma_G} a + \frac{c_u \lambda}{|c_v| \sigma_G} h(a)\right\} \frac{\partial^2 f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \\
&= \left\{-\frac{1}{\sigma_G^2} + \left(\frac{c_u \lambda}{|c_v| \sigma_G}\right)^2 [a h(a) - h(a)^2]\right\} \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \\
&\quad - \left\{\frac{|c_v|}{c_u \lambda \sigma_G} a + \frac{c_u \lambda}{|c_v| \sigma_G} h(a)\right\} \frac{\partial^2 f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T},
\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial \lambda^2} &= \frac{1}{\lambda^2} a^2 h(a)[a - h(a)] \\ &= \frac{1}{\lambda^2}[a^3 h(a) - a^2 h(a)^2],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial (\sigma_G^2)^2} &= \frac{1}{2\sigma_G^4} - \left(\frac{|c_v|}{c_u \lambda \sigma_G^2}\right)^2 a^2 - \frac{3}{4\sigma_G^4} a h(a) + \frac{1}{4\sigma_G^4} a^2 h(a)[a - h(a)] \\ &= \frac{1}{2\sigma_G^4} - \left(\frac{|c_v|}{c_u \lambda \sigma_G^2}\right)^2 a^2 - \frac{1}{4\sigma_G^4}\left[3ah(a) - a^3 h(a) + a^2 h(a)^2\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \lambda} &= -\left\{\frac{c_u}{|c_v|\sigma_G} h(a) - \frac{c_u}{|c_v|\sigma_G} a h(a)[a - h(a)]\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= -\frac{c_u}{|c_v|\sigma_G}\left\{h(a) - a^2 h(a) + a h(a)^2\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \sigma_G^2} &= -\left\{-\frac{|c_v|}{c_u \lambda \sigma_G^3} a - \frac{c_u \lambda}{2|c_v|\sigma_G^3} h(a) + \frac{c_u \lambda}{2|c_v|\sigma_G^3} a h(a)[a - h(a)]\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \left\{\frac{|c_v|}{c_u \lambda \sigma_G^3} a + \frac{c_u \lambda}{2|c_v|\sigma_G^3}[h(a) - a^2 h(a) + a h(a)^2]\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial \lambda \partial \sigma_G^2} &= \frac{1}{2\lambda\sigma_G^2} a h(a) - \frac{1}{2\lambda\sigma_G^2} a^2 h(a)[a - h(a)] \\ &= \frac{1}{2\lambda\sigma_G^2}[ah(a) - a^3 h(a) + a^2 h(a)^2].\end{aligned}$$

$$I_i(\boldsymbol{\theta}) = \mathbb{E}\left[\begin{array}{c|cc} \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right) \\ \hline \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)^2 & \left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right) \\ \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)^2 \end{array}\right]. \quad (2.12)$$

Figure 2.3: Partitioned per observation expected Fisher information matrix for the normal-half normal model.

2.2.2 Information matrix in terms of first-order partial derivatives

Appendix D provides background information on information matrices. For random variable Y_i, equation (D.2) gives the per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \sigma_G)$ as

$$I_i(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right].$$

When the parameter vector $\boldsymbol{\theta}$ is partitioned such that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$, the formula for the partitioned information matrix is given by equation (D.4) in Appendix D. Equation (2.12) in Figure 2.3 shows the form of the partitioned information matrix when $\boldsymbol{\tau} = (\lambda, \sigma_G)$. This formulation uses the first-order partial derivatives of $\ln f_{Y_i}$. Dispensing with the observation subscripts, the components of the per observation expected Fisher information matrix are

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)^T\right] = \left\{\left(\frac{|c_v|}{c_u \lambda \sigma_G}\right)^2 \mathbb{E}[a^2] + \frac{2}{\sigma_G^2}\mathbb{E}[ah(a)] + \left(\frac{c_u \lambda}{|c_v|\sigma_G}\right)^2 \mathbb{E}[h(a)^2]\right\}\left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T,$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)^2\right] = \frac{1}{\lambda^2}\mathbb{E}[a^2 h(a)^2],$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)^2\right] = \left(\frac{1}{2\sigma_G^2}\right)^2 - \frac{1}{2}\left(\frac{|c_v|}{c_u \lambda \sigma_G^2}\right)^2 \mathbb{E}[a^2] - \frac{1}{2}\left(\frac{1}{\sigma_G^2}\right)^2 \mathbb{E}[ah(a)] + \frac{1}{4}\left(\frac{|c_v|}{c_u \lambda \sigma_G}\right)^4 \mathbb{E}[a^4] + \frac{1}{2}\left(\frac{|c_v|}{c_u \lambda \sigma_G^2}\right)^2 \mathbb{E}[a^3 h(a)] + \left(\frac{1}{2\sigma_G^2}\right)^2 \mathbb{E}[a^2 h(a)^2],$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\right] = \frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \times \left\{\frac{|c_v|}{c_u\lambda^2\sigma_G}\mathbb{E}[a^2h(a)] + \frac{c_u}{|c_v|\sigma_G}\mathbb{E}[ah(a)^2]\right\},$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] = -\left\{-\frac{|c_v|}{2c_u\lambda\sigma_G^3}\mathbb{E}[a] + \frac{1}{2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^3\mathbb{E}[a^3] + \frac{|c_v|}{2c_u\lambda\sigma_G^3}\mathbb{E}[a^2h(a)] - \frac{c_u\lambda}{2|c_v|\sigma_G^3}\mathbb{E}[h(a)] + \frac{|c_v|}{2c_u\lambda\sigma_G^3}\mathbb{E}[a^2h(a)] + \frac{c_u\lambda}{2|c_v|\sigma_G^3}\mathbb{E}[ah(a)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] = \frac{1}{2\lambda\sigma_G^2}\mathbb{E}[ah(a)] - \frac{1}{2\lambda}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a^3h(a)] - \frac{1}{2\lambda\sigma_G^2}\mathbb{E}[a^2h(a)^2].$$

Calculation of the expected information matrix requires calculation of the expectation

$$\mathbb{E}[a^r \cdot h(a)^s] = \mathbb{E}\left[\left\{-\frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}\right\}^r \cdot \left\{\frac{\phi\left(\frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}\right)}{\Phi\left(\frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}\right)}\right\}^s\right], \qquad r, s \in \mathbb{N}_0,$$

which is a complicated integral. Section 3.2 of Chapter 3 gives an approximation for this quantity.

An alternative approach for calculating the per observation expected information matrix is to first approximate the derivatives $\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}$ by $\widehat{\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}}$. The

approximated information matrix can then be calculated as

$$I_i(\boldsymbol{\theta}) \;=\; \mathbb{E}\left[\left(\widehat{\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}}\right)\left(\widehat{\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}}\right)^T\right].$$

This approach eliminates the need to calculate or approximate $\mathbb{E}[a^r \cdot h(a)^s]$ and will ensure positive semidefiniteness of the information matrix. The details for approximating the first-order derivatives of $\ln f_Y$ with respect to $\boldsymbol{\theta}$ are given in Section 3.1 of Chapter 3.

2.2.3 Information matrix in terms of second-order partial derivatives

Equation (D.3) in Appendix D gives an alternative formulation for the per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \sigma_G)$ as

$$I_i(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right].$$

When the parameter vector $\boldsymbol{\theta}$ is partitioned such that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$, the formula for the partitioned information matrix is given by equation (D.5) in Appendix D. If $\boldsymbol{\tau} = (\lambda, \sigma_G)$ then the partitioned information matrix is

$$I_i(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\begin{array}{c|cc} \dfrac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\beta} \partial \lambda} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\beta} \partial \sigma_G^2} \\ \hline \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\beta} \partial \lambda}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial \lambda^2} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial \lambda \partial \sigma_G^2} \\ \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\beta} \partial \sigma_G^2}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial \lambda \partial \sigma_G^2}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial (\sigma_G^2)^2} \end{array}\right]. \qquad (2.13)$$

This formulation uses the second-order partial derivatives of $\ln f_{Y_i}$. Dispensing with the observation subscripts, the components of the per observation expected Fisher information matrix are

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] &= \\
&-\left\{-\frac{1}{\sigma_G^2}+\left(\frac{c_u\lambda}{|c_v|\sigma_G}\right)^2\left(\mathbb{E}[ah(a)]-\mathbb{E}[h(a)^2]\right)\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right) \\
&+\left\{\frac{|c_v|}{c_u\lambda\sigma_G}\mathbb{E}[a]+\frac{c_u\lambda}{|c_v|\sigma_G}\mathbb{E}[h(a)]\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},
\end{aligned}
$$

$$
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\lambda^2}\right] = -\frac{1}{\lambda^2}\left(\mathbb{E}[a^3h(a)]-\mathbb{E}[a^2h(a)^2]\right),
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(\sigma_G^2)^2}\right] &= \\
&-\frac{1}{2\sigma_G^4}+\left(\frac{|c_v|}{c_u\lambda\sigma_G^2}\right)^2\mathbb{E}[a^2]+\frac{1}{4\sigma_G^4}\left(3\mathbb{E}[ah(a)]-\mathbb{E}[a^3h(a)]+\mathbb{E}[a^2h(a)^2]\right),
\end{aligned}
$$

$$
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\lambda}\right] = \frac{c_u}{|c_v|\sigma_G}\left\{\mathbb{E}[h(a)]-\mathbb{E}[a^2h(a)]+\mathbb{E}[ah(a)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_G^2}\right] &= \\
&-\left\{\frac{|c_v|}{c_u\lambda\sigma_G^3}\mathbb{E}[a]+\frac{c_u\lambda}{2|c_v|\sigma_G^3}\left(\mathbb{E}[h(a)]-\mathbb{E}[a^2h(a)]+\mathbb{E}[ah(a)^2]\right)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}
$$

$$
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\lambda\partial\sigma_G^2}\right] = -\frac{1}{2\lambda\sigma_G^2}\left(\mathbb{E}[ah(a)]-\mathbb{E}[a^3h(a)]+\mathbb{E}[a^2h(a)^2]\right).
$$

Calculation of the information matrix requires calculation of the expectation $\mathbb{E}\left[a^r\cdot h(a)^s\right]$, $r,s\in\mathbb{N}_0$. As discussed in Section 2.2.2, this is a complicated expectation to calculate. An approximation for this quantity is given in Section 3.2 of Chapter 3.

2.3 Normal-Exponential Model

Assume that random variables U and V are distributed as follows

(i) $U \sim \text{Exponential}(1/\sigma_u)$ i.i.d.

(ii) $V \sim N(0, \sigma_v^2)$ i.i.d.

(iii) U and V are distributed independently of each other.

The exponential distribution is a special case of the gamma distribution which is discussed in Section 2.5. Figure 2.4 plots three different exponential probability density functions.

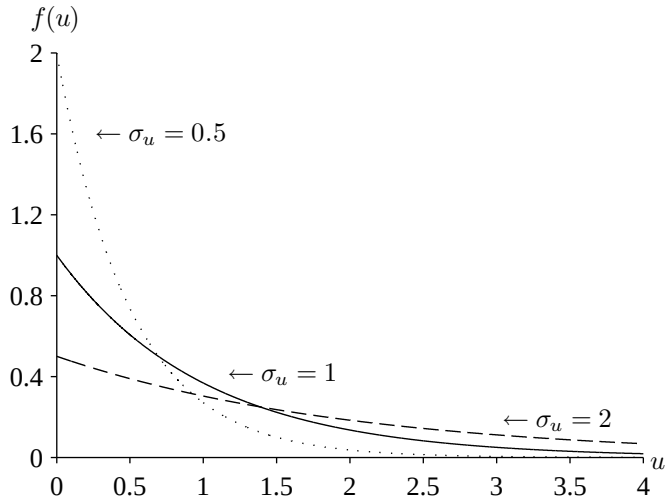

Figure 2.4: Exponential distributions.

The derivation of the information matrix for all models in this dissertation follows a similar procedure as the previous section for the normal-half normal model. The main results for the normal-exponential model are given below. The detailed calculations can be found in Appendix A.1.

The probability density function of $E = c_u U + c_v V$ is given by

$$f_E(\varepsilon) = \frac{1}{|c_u|\sigma_u} \exp\left\{-\frac{\varepsilon}{c_u\sigma_u} + \frac{c_v^2\sigma_v^2}{2c_u^2\sigma_u^2}\right\} \Phi\left(\frac{c_u\varepsilon}{|c_uc_v|\sigma_v} - \frac{|c_v|\sigma_v}{|c_u|\sigma_u}\right), \quad (2.14)$$

with mean and variance given by

$$\mathbb{E}[E] = c_u\sigma_u, \quad (2.15)$$

$$Var(E) = c_u^2\sigma_u^2 + c_v^2\sigma_v^2. \quad (2.16)$$

As with the normal-half normal model, the conditional density of U given E can be calculated using equation (C.12), but with

$$K = \frac{1}{|c_v|\sqrt{2\pi}\sigma_u\sigma_v},\ A = \frac{c_u^2}{c_v^2\sigma_v^2},\ B = \frac{c_u\varepsilon}{c_v^2\sigma_v^2} - \frac{1}{\sigma_u} \text{ and } C = \frac{\varepsilon^2}{c_v^2\sigma_v^2}.$$

The expected value and mode of U given E can be calculated using equations (C.13) and (C.14) respectively.

2.3.1 Log-likelihood function

Under model (2.2), the log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_v)$ for a sample of N independent observations can be obtained using equation (2.14) and is given by

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{1=1}^{N} \left\{\ln\left(\frac{1}{|c_u|\sigma_u}\right) - \frac{y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})}{c_u\sigma_u} + \frac{c_v^2\sigma_v^2}{2c_u^2\sigma_u^2} + \ln\Phi(-a_i)\right\}, \quad (2.17)$$

where

$$a_i = -\frac{c_u\varepsilon_i}{|c_uc_v|\sigma_v} + \frac{|c_v|\sigma_v}{|c_u|\sigma_u}.$$

Omitting observation subscripts, the expected value and variance of a, which will be used in later chapters, are

$$\mathbb{E}[a] = -\frac{c_u}{|c_uc_v|\sigma_v}\mathbb{E}[E] + \frac{|c_v|\sigma_v}{|c_u|\sigma_u},$$

$$Var(a) = \left(\frac{c_u}{|c_uc_v|\sigma_v}\right)^2 Var(E),$$

where $\mathbb{E}[E]$ and $Var(E)$ are given in equations (2.15) and (2.16) respectively. The derivative of a with respect to the parameter vector $\boldsymbol{\beta}$, which will also be used in later chapters, is

$$\frac{\partial a}{\partial \boldsymbol{\beta}} = \left(\frac{c_u}{|c_u c_v|\sigma_v}\right)\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

The first-order derivatives of $\ln f_Y(y;\boldsymbol{\theta})$ are

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= -\left\{-\frac{1}{c_u\sigma_u} + \frac{c_u}{|c_u c_v|\sigma_v}h(a)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\\
\frac{\partial \ln f_Y}{\partial(1/\sigma_u)} &= \sigma_u + \frac{|c_v|\sigma_v}{|c_u|}[a - h(a)],\\
\frac{\partial \ln f_Y}{\partial \sigma_v^2} &= \frac{c_v^2}{2c_u^2\sigma_u^2} - \left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v} - \frac{1}{2\sigma_v^2}a\right)h(a).
\end{aligned}$$

Only the first-order derivatives will be used in later chapters, hence the second-order derivatives are not given here.

2.3.2 Information matrix in terms of first-order partial derivatives

Equation (2.18) in Figure 2.5 shows the form of the partitioned information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_v)$.

The remarks that were made about approximating the information matrix for the normal-half normal model in Section 2.2 apply with equal force to all models discussed in this dissertation.

$$
I_i(\boldsymbol{\theta}) = \mathbb{E}\left[\begin{array}{c|cc}
\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right) \\
\hline
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)^2 & \left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right) \\
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)^2
\end{array}\right] \tag{2.18}
$$

Figure 2.5: Partitioned per observation expected Fisher information matrix for the normal-exponential model.

2.3.3 Information matrix in terms of second-order partial derivatives

An alternative formulation for the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_v)$ is

$$I_i(\boldsymbol{\theta}) = -\mathbb{E}\left[\begin{array}{c|cc} \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\sigma_v^2} \\ \hline \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(1/\sigma_u)^2} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(1/\sigma_u)\partial\sigma_v^2} \\ \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\sigma_v^2}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial(1/\sigma_u)\partial\sigma_v^2}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(\sigma_v^2)^2} \end{array}\right] . \quad (2.19)$$

This formulation uses the second-order partial derivatives of $\ln f_{Y_i}$, which are given in Appendix A.1.

2.4 Normal-Truncated Normal Model

Assume that random variables U and V are distributed as follows

(i) $U \sim N^+(\mu, \sigma_u^2)$ i.i.d., i.e. nonnegative truncated normal

(ii) $V \sim N(0, \sigma_v^2)$ i.i.d.

(iii) U and V are distributed independently of each other.

The nonnegative truncated normal distribution considered here is the normal distribution, with mean $\mu \in \mathbb{R}$, which is truncated from below at zero. When $\mu = 0$ the nonnegative truncated normal distribution simplifies to the nonnegative half normal distribution of Section 2.2. Appendix C.5 provides further details on

truncated normal distributions. Three different truncated normal distributions are plotted in Figure 2.6 where $\sigma_u = 1$ for all densities and μ is negative, zero (the half normal case) and positive.

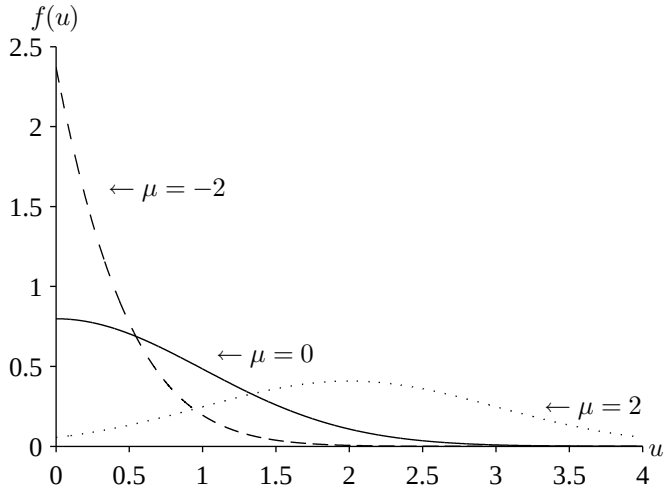

Figure 2.6: Truncated normal distributions with $\sigma_u = 1$.

As with the normal-half normal model, if we let

$$\sigma_G^2 = c_u^2\sigma_u^2 + c_v^2\sigma_v^2, \tag{2.20}$$

$$\lambda = \frac{\sigma_u}{\sigma_v}, \tag{2.21}$$

then the probability density function of $E = c_u U + c_v V$ is given by

$$f_E(\varepsilon) = \frac{1}{\sigma_G}\phi\left(\frac{c_u\mu - \varepsilon}{\sigma_G}\right)\Phi\left(\frac{|c_v|\mu}{\lambda\sigma_G} + \frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}, \tag{2.22}$$

with mean and variance given by

$$\mathbb{E}[E] = \tilde{c_u}\sigma_u, \tag{2.23}$$

$$Var(E) = \tilde{\tilde{c_u}}^{\,2}\sigma_u^2 + c_v^2\sigma_v^2, \tag{2.24}$$

where $\tilde{c_u} = \dfrac{c_u\mu}{\sigma_u} + c_u h\left(-\dfrac{\mu}{\sigma_u}\right)$ and $\tilde{\tilde{c_u}}^{\,2} = c_u^2\left\{1 - \dfrac{\mu}{\sigma_u}h\left(-\dfrac{\mu}{\sigma_u}\right) - \left[h\left(-\dfrac{\mu}{\sigma_u}\right)\right]^2\right\}$.

The conditional density of U given E can be calculated using equation (C.12) with

$$K = \frac{1}{|c_v|2\pi\sigma_u\sigma_v}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1},\ A = \frac{1}{\sigma_u^2} + \frac{c_u^2}{c_v^2\sigma_v^2},\ B = \frac{\mu}{\sigma_u^2} + \frac{c_u\varepsilon}{c_v^2\sigma_v^2} \text{ and}$$
$$C = \frac{\mu^2}{\sigma_u^2} + \frac{\varepsilon^2}{c_v^2\sigma_v^2}.$$

The expected value and mode of U given E can be calculated using equations (C.13) and (C.14) respectively.

2.4.1 Log-likelihood function

Under model (2.2), the log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \lambda, \sigma_G)$ for a sample of N independent observations can be obtained using equation (2.22) and is given by

$$\begin{aligned}\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{i=1}^{N}\left\{-\ln\sigma_G + \ln\phi\left(\frac{c_u\mu - [y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})]}{\sigma_G}\right) + \ln\Phi(-a_{1i})\right. \\ &\qquad \left. - \ln\Phi\left(-a_2\right)\right\}, \end{aligned} \tag{2.25}$$

where

$$a_{1i} = -\frac{|c_v|\mu}{\lambda\sigma_G} - \frac{c_u\lambda\varepsilon_i}{|c_v|\sigma_G}.$$

The parameter σ_u can be expressed as a function of λ and σ_G by solving (2.20) and (2.21) simultaneously to give

$$\sigma_u = \frac{\lambda\sigma_G}{(c_u^2\lambda^2 + c_v^2)^{1/2}},$$

which can be substituted into the formula for a_2 to give

$$\begin{aligned} a_2 &= -\frac{\mu}{\sigma_u} \\ &= -\frac{\mu(c_u^2\lambda^2 + c_v^2)^{1/2}}{\lambda\sigma_G}. \end{aligned}$$

Omitting observation subscripts, the expected value and variance of a_1, which will be used in later chapters, are

$$\begin{aligned}\mathbb{E}[a_1] &= -\frac{|c_v|\mu}{\lambda\sigma_G} - \frac{c_u\lambda}{|c_v|\sigma_G}\mathbb{E}[E],\\ Var(a_1) &= \left(\frac{c_u\lambda}{|c_v|\sigma_G}\right)^2 Var(E),\end{aligned}$$

where $\mathbb{E}[E]$ and $Var(E)$ are given in equations (2.23) and (2.24) respectively. The derivative of a_1 with respect to the parameter vector $\boldsymbol{\beta}$, which will also be used in later chapters, is

$$\frac{\partial a_1}{\partial \boldsymbol{\beta}} = \left(\frac{c_u\lambda}{|c_v|\sigma_G}\right)\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

The first-order derivatives of $\ln f_Y(y;\boldsymbol{\theta})$ are

$$\begin{aligned}\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= -\left\{\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2} + \frac{|c_v|}{c_u\lambda\sigma_G}a_1 + \frac{c_u\lambda}{|c_v|\sigma_G}h(a_1)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\\ \frac{\partial \ln f_Y}{\partial \mu} &= -\frac{\mu(c_u^2\lambda^2+c_v^2)}{\lambda^2\sigma_G^2} - \frac{|c_v|}{\lambda\sigma_G}a_1 + \frac{|c_v|}{\lambda\sigma_G}h(a_1) + \frac{a_2}{\mu}h(a_2),\\ \frac{\partial \ln f_Y}{\partial \lambda} &= \left(-\frac{2|c_v|\mu}{\lambda^2\sigma_G} - \frac{1}{\lambda}a_1\right)h(a_1) + \frac{c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}h(a_2),\\ \frac{\partial \ln f_Y}{\partial \sigma_G^2} &= -\frac{1}{2\sigma_G^2} + \frac{1}{2}\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2} + \frac{|c_v|}{c_u\lambda\sigma_G}a_1\right)^2 + \frac{1}{2\sigma_G^2}a_1h(a_1) - \frac{1}{2\sigma_G^2}a_2h(a_2),\end{aligned}$$

2.4.2 Information matrix in terms of first-order partial derivatives

Equation (2.26) in Figure 2.7 shows the form of the partitioned information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \lambda, \sigma_G)$.

$$
I_i(\boldsymbol{\theta}) = \mathbb{E}\left[\begin{array}{c|ccc}
\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right) \\
\hline
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right)^2 & \left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right) \\
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)^2 & \left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right) \\
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \mu}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \lambda}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_G^2}\right)^2
\end{array}\right]. \tag{2.26}
$$

Figure 2.7: Partitioned per observation expected Fisher information matrix for the normal-truncated normal model.

2.4.3 Information matrix in terms of second-order partial derivatives

An alternative formulation for the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \lambda, \sigma_G)$ is

$$I_i(\boldsymbol{\theta}) = -\mathbb{E}\left[\begin{array}{c|ccc} \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\mu} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\lambda} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\sigma_G^2} \\ \hline \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\mu}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\mu^2} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\mu\partial\lambda} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\mu\partial\sigma_G^2} \\ \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\lambda}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\mu\partial\lambda}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\lambda^2} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\lambda\partial\sigma_G^2} \\ \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\sigma_G^2}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\mu\partial\sigma_G^2}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\lambda\partial\sigma_G^2}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(\sigma_G^2)^2} \end{array}\right]. \tag{2.27}$$

This formulation uses the second-order partial derivatives of $\ln f_{Y_i}$, which are given in Appendix A.2.

2.5 Normal-Gamma Model

Assume that random variables U and V are distributed as follows

(i) $U \sim \text{Gamma}(\alpha, \sigma_u)$ i.i.d.

(ii) $V \sim N(0, \sigma_v^2)$ i.i.d.

(iii) U and V are distributed independently of each other.

When $\alpha = 1$ the gamma distribution simplifies to the exponential distribution of Section 2.3. Three different gamma distributions are plotted in Figure 2.8 where $\sigma_u = 1$ for all densities and $\alpha = 1, 2, 3$. When $0 < \alpha < 1$ the gamma density looks like an exponential density while $\alpha > 1$ has a mode farther away from zero as α increases.

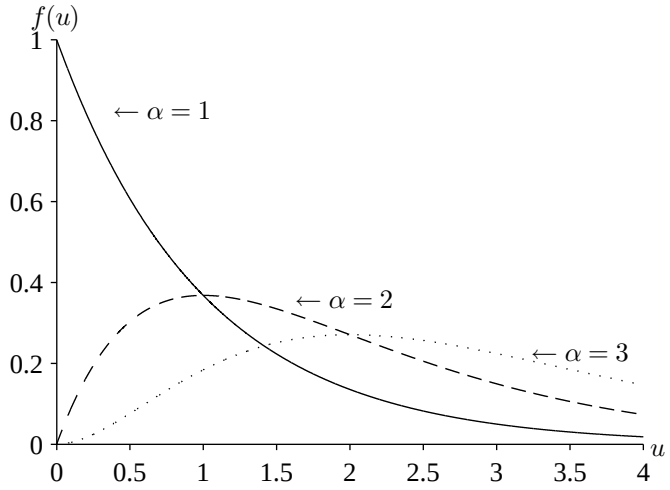

Figure 2.8: Gamma distributions with $\sigma_u = 1$.

The marginal density of $E = c_u U + c_v V$ is given by

$$f_E(\varepsilon) = \frac{1}{|c_u|\Gamma(\alpha)\sigma_u^\alpha} \exp\left\{-\frac{\varepsilon}{c_u\sigma_u} + \frac{c_v^2\sigma_v^2}{2c_u^2\sigma_u^2}\right\} \Phi\left(\frac{c_u\varepsilon}{|c_u c_v|\sigma_v} - \frac{|c_v|\sigma_v}{|c_u|\sigma_u}\right) \mathbb{E}[Q^{\alpha-1}], \tag{2.28}$$

where $\mathbb{E}[Q^{\alpha-1}]$ is a fractional moment of the nonnegative truncated normal distribution of random variable Q. The mean and variance of E are

$$\mathbb{E}[E] = \tilde{c_u}\sigma_u, \tag{2.29}$$

$$Var(E) = \tilde{\tilde{c_u}}^2\sigma_u^2 + c_v^2\sigma_v^2, \tag{2.30}$$

where $\tilde{c_u} = c_u\alpha$ and $\tilde{\tilde{c_u}}^2 = c_u^2\alpha$.

The conditional density of U given E can be calculated using equation (C.9) and is given by

$$f_{U|E}(u|\varepsilon) = \frac{u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)}{\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du} = \frac{u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)}{\Phi\left(\frac{B}{\sqrt{A}}\right)\mathbb{E}[Q^{\alpha-1}]},$$

with expected value given by equation (C.10) as

$$\mathbb{E}[U|E] = \frac{\int_0^\infty u^{\alpha}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du}{\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du} = \frac{\mathbb{E}[Q^{\alpha}]}{\mathbb{E}[Q^{\alpha-1}]},$$

and where

$$K = \frac{1}{|c_v|\Gamma(\alpha)\sigma_u^\alpha\sqrt{2\pi}\sigma_v},\ A = \frac{c_u^2}{c_v^2\sigma_v^2},\ B = \frac{c_u\varepsilon}{c_v^2\sigma_v^2} - \frac{1}{\sigma_u} \text{ and } C = \frac{\varepsilon^2}{c_v^2\sigma_v^2}.$$

When $c_u = -1$ and $c_v = 1$, the marginal density of E given in equation (2.28) is the marginal density function derived by Greene (1990). Since α need not be an integer, there is no closed form for $\mathbb{E}[Q^{\alpha-1}]$ and hence no closed form for the density of E. Consequently, approximation methods must be employed in evaluating the marginal density of E and its log-likelihood function.

Beckers & Hammond (1987) derived a closed form expression for the marginal density of E, when $c_u = 1$ and $c_v = 1$, which does not restrict α to integer values. Although their formulation is appealing because the marginal density of E and its log-likelihood function can be evaluated analytically, it shall not be considered here due to practical considerations. Beckers & Hammond advise that their approach is complex and impractical if interest is in evaluating the Hessian matrix of the log-likelihood function. Additionally, approximation methods are likely to be needed in calculating the information matrix thus negating the benefits of the analytical formulation.

Nakamura (1980) discusses moments of positively truncated normal distributions. Nakamura's approximation of $\mathbb{E}[Q^{\alpha-1}]$ restricts α to be an integer less than or equal to zero. However when α originates as a parameter from the gamma distribution, α is strictly positive.

2.5.1 Log-likelihood function

Under model (2.2), the log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$ for a sample of N independent observations can be obtained using equation (2.28) and is given by

$$\begin{aligned}\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) &= \sum_{i=1}^{N} \left\{ -\ln\left(|c_u|\Gamma(\alpha)\right) + \alpha \ln\left(\frac{1}{\sigma_u}\right) - \frac{y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})}{c_u \sigma_u} + \frac{c_v^2 \sigma_v^2}{2 c_u^2 \sigma_u^2} \right. \\ &\qquad \left. + \ln\left(\frac{|c_u|}{|c_v|\sigma_v}\right) + \ln\left(\int_0^\infty u_i^{\alpha-1} \phi(-a_i)\, du_i\right) \right\}, \qquad (2.31)\end{aligned}$$

where

$$a_i = -\frac{|c_u|}{|c_v|\sigma_v} u_i + \frac{c_u \varepsilon_i}{|c_u c_v|\sigma_v} - \frac{|c_v|\sigma_v}{|c_u|\sigma_u}.$$

Omitting observation subscripts, the derivative of $\varepsilon = y - f(\boldsymbol{x}, \boldsymbol{\beta})$ with respect to the parameter vector $\boldsymbol{\beta}$, which will be used in later chapters, is

$$\frac{\partial \varepsilon}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}}[y - f(\boldsymbol{x}, \boldsymbol{\beta})] = -\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

The first-order derivatives of $\ln f_Y(y; \boldsymbol{\theta})$ with respect to the parameters of interest are

$$\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} = \frac{1}{c_v^2 \sigma_v^2} \left\{\varepsilon - c_u \mathbb{E}[U|E]\right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

$$\frac{\partial \ln f_Y}{\partial \alpha} = -\psi(\alpha) + \ln\left(\frac{1}{\sigma_u}\right) + \mathbb{E}[\ln U|E],$$

$$\frac{\partial \ln f_Y}{\partial (1/\sigma_u)} = \alpha \sigma_u - \mathbb{E}[U|E],$$

$$\frac{\partial \ln f_Y}{\partial \sigma_v^2} = -\frac{1}{2\sigma_v^2} + \frac{1}{2c_v^2\sigma_v^4}\left(\varepsilon^2 + c_u^2\mathbb{E}[U^2|E] - 2c_u\varepsilon\mathbb{E}[U|E]\right),$$

where

$$\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{d\ln\Gamma(\alpha)}{d\alpha},$$

is the digamma function and equation (C.11) in Appendix C gives

$$\mathbb{E}[g(U)|E] = \frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}.$$

2.5.2 Information matrix in terms of first-order partial derivatives

Equation (2.33) in Figure 2.9 shows the form of the partitioned information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$. Calculation of the expected information matrix requires the evaluation of complicated integrals. The integrals appear in the derivatives through the conditional expectation $\mathbb{E}[g(U)|E]$. Although the integrals can be approximated numerically using some form of Gaussian quadrature, care should be taken in the choice of quadrature rule employed. Abramowitz & Stegun (1965) give various quadrature rules. Because the integrals in $\mathbb{E}[g(U)|E]$ are over the interval $[0, \infty)$, the Gauss-Laguerre formula is one such quadrature rule that can be applied. However several methods should be implemented so that the sensitivity of the values in the information matrix to the method of numerical integration used can be assessed.

As with the model specifications from previous sections of this chapter, an alternative approach for calculating the per observation expected information matrix is to first approximate the derivatives $\dfrac{\partial \ln f_Y}{\partial \boldsymbol{\theta}}$ by $\dfrac{\widehat{\partial \ln f_Y}}{\partial \boldsymbol{\theta}}$. Unlike in previous sections, this approach does not eliminate the need to calculate or approximate the expectations appearing in the information matrix. However it does simplify

calculations somewhat and will ensure positive semidefiniteness of the information matrix. Numerical integration is still required and, as before, the resultant information matrix may be sensitive to the quadrature technique employed. The details for approximating the first-order derivatives of $\ln f_Y$ with respect to $\boldsymbol{\theta}$ are given in Section 3.1 of Chapter 3.

2.5.3 Information matrix in terms of second-order partial derivatives

An alternative formulation for the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$ is

$$I_i(\boldsymbol{\theta}) = -\mathbb{E}\left[\begin{array}{c|ccc} \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\alpha} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\sigma_v^2} \\ \hline \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\alpha}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\alpha^2} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\alpha\partial(1/\sigma_u)} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\alpha\partial\sigma_v^2} \\ \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\alpha\partial(1/\sigma_u)}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(1/\sigma_u)^2} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(1/\sigma_u)\partial\sigma_v^2} \\ \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\sigma_v^2}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\alpha\partial\sigma_v^2}\right)^T & \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial(1/\sigma_u)\partial\sigma_v^2}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial(\sigma_v^2)^2} \end{array}\right]. \tag{2.32}$$

This formulation uses the second-order partial derivatives of $\ln f_{Y_i}$, which are given in Appendix A.3. As discussed in Section 2.5.2, calculation of the information matrix requires evaluating complicated integrals. The integrals appear in the derivatives through the conditional expectations, variances and covariances. Numerical integration can be utilised to approximate these integrals although care should be taken in determining the choice of quadrature method.

$$
I_i(\boldsymbol{\theta}) = \mathbb{E}\left[\begin{array}{c|ccc}
\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right) \\
\hline
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right)^2 & \left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right) & \left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right) \\
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)^2 & \left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right) \\
\left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \alpha}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)\right\}^T & \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial (1/\sigma_u)}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \sigma_v^2}\right)^2
\end{array}\right]. \tag{2.33}
$$

Figure 2.9: Partitioned per observation expected Fisher information matrix for the normal-gamma model.

Chapter 3

Approximation Methods for Information Matrices

Per observation expected Fisher information matrices were derived in Chapter 2 for the four different model specifications given in Sections 2.2 to 2.5. Under each model specification, information matrices were derived using the first-order and second-order partial derivatives of $\ln f_{Y_i}(y_i;\boldsymbol{\theta})$, where $\ln f_{Y_i}(y_i;\boldsymbol{\theta})$ is the log-likelihood function for the i-th observation. It was noted throughout Chapter 2 that the information matrices for the four model specifications should be approximated. This is because approximation will ease the evaluation of complicated expectations and integrals appearing in the information matrices.

The formula for the information matrix based on the first-order partial derivatives of $\ln f_{Y_i}(y_i;\boldsymbol{\theta})$ is defined in equation (D.2) of Appendix D and is given by

$$I_i(\boldsymbol{\theta}) = Cov\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right), \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right] = \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right], \tag{3.1}$$

where $\ln f_{Y_i} = \ln f_{Y_i}(y_i;\boldsymbol{\theta})$. The information matrix based on the second-order derivatives of $\ln f_{Y_i}(y_i;\boldsymbol{\theta})$ is defined in equation (D.3) of Appendix D and is given

by

$$I_i(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]. \tag{3.2}$$

Under certain regularity conditions, if the information matrix can be evaluated exactly, the information matrix derived using the first-order derivatives is equivalent to the information matrix derived using the second-order derivatives, that is

$$\mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right] \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]. \tag{3.3}$$

Therefore, if the information matrix can be evaluated exactly, equations (3.1) and (3.2) will produce the same information matrix.

If the information matrix cannot be evaluated exactly and is approximated, equivalence (3.3) may not hold. Consequently, if an approximation method is used in evaluating the information matrix, equations (3.1) and (3.2) will result in possibly numerically different information matrices. Additionally, the type of approximation method used can lead to numerically different information matrices. Clearly, the choice between equations (3.1) and (3.2) in approximating the information matrix is worthy of discussion and hence is the topic of this chapter.

A good approximation will produce an approximated information matrix $\widehat{I}_i(\boldsymbol{\theta})$ with values close to the true information matrix $I_i(\boldsymbol{\theta})$, so that $\widehat{I}_i(\boldsymbol{\theta}) \approx I_i(\boldsymbol{\theta})$. However, if the true information matrix cannot be evaluated then there will be no way of determining how good the approximation is to the true matrix. A sensible approach may be to compare different approximations. If the approximations are reasonably accurate then the information matrices produced by the different approximation methods should be close. Unfortunately, if the approximations are all equally bad then they may be close to each other but not to the true information matrix.

First-order Taylor series approximations are used throughout this chapter. It may be of interest to investigate the effects of higher-order approximations or

alternative approximation methods on information matrices, however they are not discussed here.

3.1 Approximating the Information Matrix of First-order Derivatives

In this section, the formula for the information matrix that will be utilised involves the first-order partial derivatives of $\ln f_{Y_i}(y_i;\boldsymbol{\theta})$. Equation (3.1) gives the form of the per observation expected Fisher information matrix as

$$I_i(\boldsymbol{\theta}) \;=\; Cov\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right),\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right] \;=\; \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right].$$

The two methods for approximating the above information matrix use a first-order Taylor series approximation. The first method approximates the first-order derivatives separately whilst the second method approximates the product of the first-order derivatives.

3.1.1 Method 1: Approximating the first-order derivatives (Recommended)

Consider the statistical model $Y_i = f(\boldsymbol{x}_i,\boldsymbol{\beta}) + E_i$ with probability density function $f_{Y_i} = f_{Y_i}(y_i;\boldsymbol{\theta})$ where the k-dimensional parameter vector $\boldsymbol{\theta}$ is partitioned such that $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\tau})$. The $\boldsymbol{\beta}$ parameters originate from the model and the $\boldsymbol{\tau}$ parameters arise from the distributional assumption on Y_i.

Let a_i be a function of random variable Y_i. Using equation (3.1) as the definition of an information matrix, the per observation expected Fisher information matrix can be written as

$$I_i(\boldsymbol{\theta}) \;=\; Cov\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right),\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right] \;=\; Cov\left[\boldsymbol{f}_\theta(a_i,\boldsymbol{x}_i),\boldsymbol{f}_\theta^T(a_i,\boldsymbol{x}_i)\right],$$

where

$$\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}} = \boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i).$$

That is, the first-order derivatives are functions of a_i and $\boldsymbol{x}_i$. The following method approximates the information matrix for the i-th observation by approximating the functions $\boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i)$ using a first-order Taylor polynomial.

Let $\mu_a = \mathbb{E}[a_i]$ and $\sigma_a^2 = Var(a_i)$. The first-order Taylor series approximation of $\boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i)$ about $a_i = \mu_a$ can be derived using equation (C.20) in Appendix C.7 and is given by

$$\widehat{\boldsymbol{f}}_\theta(a_i, \boldsymbol{x}_i) = \boldsymbol{f}_\theta(\mu_a, \boldsymbol{x}_i) + (a_i - \mu_a)\, \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i),$$

where

$$\boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) = \left.\frac{\partial \boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i)}{\partial a_i}\right|_{a_i=\mu_a} = \left.\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial a_i}\right|_{a_i=\mu_a}.$$

Thus the approximated per observation expected Fisher information matrix for the i-th observation is

$$\begin{aligned}
\widehat{I}_i(\boldsymbol{\theta}) &= Cov\left[\widehat{\boldsymbol{f}}_\theta(a_i, \boldsymbol{x}_i), \widehat{\boldsymbol{f}}_\theta^T(a_i, \boldsymbol{x}_i)\right] \\
&= Cov\left[(a_i - \mu_a)\, \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i), (a_i - \mu_a)\, \boldsymbol{f}'^T_\theta(\mu_a, \boldsymbol{x}_i)\right] \\
&= \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\, \boldsymbol{f}'^T_\theta(\mu_a, \boldsymbol{x}_i)\, Cov\left[(a_i - \mu_a), (a_i - \mu_a)\right] \\
&= \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\, \boldsymbol{f}'^T_\theta(\mu_a, \boldsymbol{x}_i)\, \sigma_a^2. \qquad (3.4)
\end{aligned}$$

The advantage of the form of this approximated per observation information matrix is that it is positive semidefinite. As a result, the full information matrix for all N observations will be either positive definite or positive semidefinite. If the full information matrix is positive definite, it will be nonsingular and all the parameters will be estimable. If the full information matrix is positive semidefinite, it will be singular, however subsets or linear combinations of the parameters will be estimable.

Calculating the approximated information matrix

Consider the per observation expected Fisher information matrices derived in Chapter 2. In Sections 2.2 and 2.3, a_i is a function of parameter vector $\boldsymbol{\beta}$, hence the chain rule can be used to calculate $\boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)$ as

$$\boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) = \frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}^T} \cdot \left(\frac{\partial a_i}{\partial \boldsymbol{\beta}} \right)^{-1} \Bigg|_{a_i = \mu_a}$$

where $\dfrac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}^T} = \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}^T} \right)^T$ and $\dfrac{\partial a_i}{\partial \boldsymbol{\beta}}$ were derived in Sections 2.2 and 2.3.

The derivatives of $\ln f_{Y_i}(y_i; \boldsymbol{\theta})$ in Section 2.4 are given as functions of a_{1i} rather than functions of a_i. This is just a notational difference. The approximation of the information matrix under the model specification given in Section 2.4 can be derived in the same manner as detailed above by simply substituting a_{1i} for a_i into the above equations. Similarly, the approximated information matrix under the model specification given in Section 2.5 can be derived by substituting ε_i for a_i into the above equations.

Properties of the approximated information matrix

Bhatia (2007) provides details on the properties of positive definite matrices. Positive semidefiniteness of the approximated per observation information matrix and the approximated full information matrix are established in the following theorems.

Theorem 3.1.1 The approximated per observation expected Fisher information matrix

$$\widehat{I}_i(\boldsymbol{\theta}) = \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\, \boldsymbol{f}'^T_\theta(\mu_a, \boldsymbol{x}_i)\, \sigma_a^2,$$

given in equation (3.4) is positive semidefinite with rank $\widehat{I}_i(\boldsymbol{\theta}) \leq 1$.

Proof To simplify notation, let $\boldsymbol{f} = \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)$ with the j-th element denoted by f_j, $j = 1, \ldots, k$. For all non-zero vectors $\boldsymbol{v} \in \mathbb{R}^k$,

$$\begin{aligned} \boldsymbol{v}^T \boldsymbol{f}\boldsymbol{f}^T \boldsymbol{v} &= \left(\boldsymbol{v}^T \boldsymbol{f}\right)\left(\boldsymbol{v}^T \boldsymbol{f}\right)^T \\ &= \left(\boldsymbol{v}^T \boldsymbol{f}\right)^2 \\ &\geq 0. \end{aligned}$$

Thus the matrix $\boldsymbol{f}\boldsymbol{f}^T$ is positive semidefinite. Consequently, $\widehat{I}_i(\boldsymbol{\theta}) = \boldsymbol{f}\boldsymbol{f}^T\sigma_a^2$ is positive semidefinite.

To prove the statement about the rank, we use a standard result from linear algebra. Note that $\boldsymbol{f}\boldsymbol{f}^T$ can be expressed as

$$\boldsymbol{f}\boldsymbol{f}^T = \boldsymbol{f} \cdot [f_1, f_2, \ldots, f_k].$$

Clearly the k column vectors $\boldsymbol{f} \cdot f_j$ are not linearly independent as they are just proportional to the column vector $\boldsymbol{f}$, that is

$$\boldsymbol{f} \cdot f_j \propto \boldsymbol{f}, \qquad j = 1, \ldots, k.$$

Hence

$$\begin{aligned} \text{rank}\left\{\widehat{I}_i(\boldsymbol{\theta})\right\} &= \text{rank}\left\{\boldsymbol{f}\boldsymbol{f}^T\sigma_a^2\right\} \\ &= \text{rank}\left\{\boldsymbol{f}\boldsymbol{f}^T\right\} \\ &\leq 1. \end{aligned}$$

The rank will be zero if $\boldsymbol{f}$ is the null vector. □

Theorem 3.1.2 Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ be the parameter vector associated with covariates $\boldsymbol{x}_i$ through the model $Y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + E_i$ and let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-p})$ be the parameter vector arising from the distributional assumption on Y_i.

If the k-dimensional parameter vector $\boldsymbol{\theta}$ is partitioned such that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ then the approximated expected Fisher information matrix, weighted per obser-

vation, given by

$$\widehat{I}(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} w_i \widehat{I}_i(\boldsymbol{\theta}),$$

is positive semidefinite with

$$\text{rank } \widehat{I}(\boldsymbol{\theta}) \;\leq\; p+1,$$

where $0 \leq w_i \leq 1$, $\sum_{i=1}^{n} w_i = 1$.

Proof The sum of positive semidefinite matrices is positive semidefinite. From Theorem 3.1.1, $\widehat{I}_i(\boldsymbol{\theta})$ is positive semidefinite therefore

$$\widehat{I}(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{n} w_i \widehat{I}_i(\boldsymbol{\theta}),$$

is positive semidefinite.

To prove the statement about the rank, first note that the rank of the sum of positive semidefinite matrices is less than or equal to the sum of the rank of each matrix. Also the rank of a matrix is less than or equal to the smallest dimension of that matrix.

For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, the first-order partial derivatives of $\ln f_{Y_i}(y_i; \boldsymbol{\theta})$ are calculated using the chain rule and, for the four error specifications considered in this thesis, can be expressed as functions of a_i and $\boldsymbol{x}_i$, that is

$$\begin{aligned}
\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}} &= \left(\frac{\partial \ln f_{Y_i}}{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}\right) \times \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) \\
&= g(a_i) \times \boldsymbol{g}(\boldsymbol{x}_i) \\
&= \boldsymbol{f}_\beta(a_i, \boldsymbol{x}_i).
\end{aligned}$$

The first-order derivatives with respect to $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-p})$ are functions of a_i alone (and not $\boldsymbol{x}_i$), that is

$$\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}} \;=\; \boldsymbol{f}_\tau(a_i).$$

Hence the first-order partial derivatives with respect to the full parameter vector can be written as

$$\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_\beta(a_i, \boldsymbol{x}_i) \\ \boldsymbol{f}_\tau(a_i) \end{bmatrix} = \boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i).$$

The derivative $\boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)$ is given by

$$\boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) = \begin{bmatrix} \boldsymbol{f}'_\beta(\mu_a, \boldsymbol{x}_i) \\ \boldsymbol{f}'_\tau(\mu_a) \end{bmatrix},$$

where

$$\boldsymbol{f}'_\beta(\mu_a, \boldsymbol{x}_i) = \left.\frac{\partial \boldsymbol{f}_\beta(a_i, \boldsymbol{x}_i)}{\partial a_i}\right|_{a_i=\mu_a} = \left.\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\beta} \partial a_i}\right|_{a_i=\mu_a},$$

$$\boldsymbol{f}'_\tau(\mu_a) = \left.\frac{\partial \boldsymbol{f}_\tau(a_i)}{\partial a_i}\right|_{a_i=\mu_a} = \left.\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\tau} \partial a_i}\right|_{a_i=\mu_a}.$$

Using equation (3.4), the approximated per observation information matrix is

$$\begin{aligned} \widehat{I}_i(\boldsymbol{\theta}) &= \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\, \boldsymbol{f}'^T_\theta(\mu_a, \boldsymbol{x}_i)\, \sigma_a^2 \\ &= \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) \left[\boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i) \;\middle|\; \boldsymbol{f}'^T_\tau(\mu_a)\right] \sigma_a^2 \\ &= \left[\boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\, \boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i) \;\middle|\; \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\, \boldsymbol{f}'^T_\tau(\mu_a)\right] \sigma_a^2, \end{aligned}$$

giving

$$\sum_{i=1}^n w_i \widehat{I}_i(\boldsymbol{\theta}) = \left[\sum_{i=1}^n w_i \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) \boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i) \;\middle|\; \left(\sum_{i=1}^n w_i \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\right) \boldsymbol{f}'^T_\tau(\mu_a)\right] \sigma_a^2.$$

Clearly the last $k-p$ column vectors, the right partition of the matrix above, are not linearly independent. The elements of $\boldsymbol{f}'^T_\tau(\mu_a)$ are functions of μ_a alone (and not $\boldsymbol{x}_i$), therefore the last $k-p$ columns are just proportional to the column vector $\sum_{i=1}^n w_i \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)$. Hence

$$\begin{aligned} \text{rank} \left\{\sum_{i=1}^n w_i \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) \boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i)\right\} &\leq p, \\ \text{rank} \left\{\left(\sum_{i=1}^n w_i \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i)\right) \boldsymbol{f}'^T_\tau(\mu_a)\right\} &\leq 1, \end{aligned}$$

giving

$$\text{rank } \widehat{I}(\boldsymbol{\theta}) \;=\; \text{rank } \left\{ \sum_{i=1}^{n} w_i \widehat{I}_i(\boldsymbol{\theta}) \right\} \;\leq\; p+1.$$

□

Corollary 3.1.1 Theorem 3.1.2 implies that the information matrix $\widehat{I}(\boldsymbol{\theta}) = \widehat{I}(\boldsymbol{\beta}, \boldsymbol{\tau})$ can only be of full rank if $\boldsymbol{\tau}$ has dimension one, i.e. if there is only one τ parameter. This is a necessary, but not a sufficient, condition.

If the information matrix is rank deficient, not all parameters will be estimable. The rank of the information matrix is less than or equal to $p+1$, therefore the maximum number of parameters (or linear combinations of parameters) that can be estimated is $p+1$.

Writing the approximated information matrix in a more expanded form gives

$$\begin{aligned}
\widehat{I}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} w_i \boldsymbol{f}'_\theta(\mu_a, \boldsymbol{x}_i) \boldsymbol{f}'^T_\theta(\mu_a, \boldsymbol{x}_i) \sigma_a^2 \\
&= \sum_{i=1}^{n} w_i \begin{bmatrix} \boldsymbol{f}'_\beta(\mu_a, \boldsymbol{x}_i) \\ \boldsymbol{f}'_\tau(\mu_a) \end{bmatrix} \begin{bmatrix} \boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i) & \boldsymbol{f}'^T_\tau(\mu_a) \end{bmatrix} \sigma_a^2 \\
&= \begin{bmatrix} \displaystyle\sum_{i=1}^{n} w_i \boldsymbol{f}'_\beta(\mu_a, \boldsymbol{x}_i) \boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i) & \left(\displaystyle\sum_{i=1}^{n} w_i \boldsymbol{f}'_\beta(\mu_a, \boldsymbol{x}_i) \right) \boldsymbol{f}'^T_\tau(\mu_a) \\ \boldsymbol{f}'_\tau(\mu_a) \left(\displaystyle\sum_{i=1}^{n} w_i \boldsymbol{f}'^T_\beta(\mu_a, \boldsymbol{x}_i) \right) & \boldsymbol{f}'_\tau(\mu_a) \boldsymbol{f}'^T_\tau(\mu_a). \end{bmatrix} \sigma_a^2.
\end{aligned}$$

The top left $p \times p$ partition of the matrix, associated with the $\boldsymbol{\beta}$ parameters, has rank less than or equal to p. Therefore p or less of the $\boldsymbol{\beta}$ parameters will be estimable. The bottom right $(k-p) \times (k-p)$ partition of the matrix, associated with the $\boldsymbol{\tau}$ parameters, has rank less than or equal to one. Therefore, at best, only one of the $\boldsymbol{\tau}$ parameters will be estimable.

3.1.2 Method 2: Approximating the product of first-order derivatives

Again, using equation (3.1) as the definition of an information matrix, the (j,l)-th element of the per observation expected Fisher information matrix can be written as

$$I_i(\boldsymbol{\theta})_{(j,l)} = \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \theta_j}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \theta_l}\right)\right] = \mathbb{E}\left[f_{j,l}(a_i, \boldsymbol{x}_i)\right],$$

where

$$f_{j,l}(a_i, \boldsymbol{x}_i) = \left(\frac{\partial \ln f_{Y_i}}{\partial \theta_j}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \theta_l}\right).$$

That is, the product of the first-order derivatives are functions of a_i and $\boldsymbol{x}_i$. The following method approximates the information matrix for the i-th observation by approximating the functions $f_{j,l}(a_i, \boldsymbol{x}_i)$ using a first-order Taylor polynomial.

The first-order Taylor series approximation of $f_{j,l}(a_i, \boldsymbol{x}_i)$ about $a_i = \mu_a$ and its expected value can be derived using equations (C.20) and (C.21) in Appendix C.7 and are given by

$$\begin{aligned} \widehat{f}_{j,l}(a_i, \boldsymbol{x}_i) &= f_{j,l}(\mu_a, \boldsymbol{x}_i) + (a_i - \mu_a) f'_{j,l}(\mu_a, \boldsymbol{x}_i), \\ \mathbb{E}\left[\widehat{f}_{j,l}(a_i, \boldsymbol{x}_i)\right] &= f_{j,l}(\mu_a, \boldsymbol{x}_i). \end{aligned}$$

Thus the (j,l)-th element of the approximated per observation expected Fisher information matrix for the i-th observation is

$$\widehat{I}_i(\boldsymbol{\theta})_{(j,l)} = \mathbb{E}\left[\widehat{f}_{j,l}(a_i, \boldsymbol{x}_i)\right] = f_{j,l}(\mu_a, \boldsymbol{x}_i). \tag{3.5}$$

This simply says that the approximated per observation expected Fisher information matrix can be calculated by evaluating any functions of a_i at $a_i = \mu_a$.

A drawback of the form of this approximated per observation information matrix is that, even with a higher-order Taylor approximation, it is not guaranteed

to be positive semidefinite. As a result, the weighted information matrix for n observations may not be invertible. Hence estimation of the parameters, or even subsets or linear combinations of the parameters may not be possible.

Calculating the approximated information matrix

The elements of the per observation expected Fisher information matrices derived in Sections 2.2 and 2.3 of Chapter 2 involve expectations of the form

$$\mathbb{E}[a_i^r \cdot h(a_i)^s], \qquad r, s \in \mathbb{N}_0.$$

Under the approximation given in equation (3.5), the information matrices can be calculated using

$$\mathbb{E}[a_i^r \cdot h(a_i)^s] \;\approx\; \mu_a^r \cdot h(\mu_a)^s,$$

where $\mu_a = \mathbb{E}[a_i]$ is given in Sections 2.2 and 2.3.

The elements of the information matrix in Section 2.4 are given as functions of a_{1i} rather than functions of a_i. Hence the approximated information matrix can be derived in the same fashion as detailed above by simply substituting a_{1i} for a_i.

Similarly, the approximated information matrix under the model specification of Section 2.5 can be calculated by evaluating any functions of ε_i at $\varepsilon_i = \mathbb{E}[E_i]$. However, even after the information matrix in Section 2.5 has been approximated, any remaining integrals will require further numerical approximation.

3.2 Approximating the Information Matrix of Second-order Derivatives

In this section, the formula for the information matrix that will be utilised involves the second-order partial derivatives of $\ln f_{Y_i}(y_i; \boldsymbol{\theta})$. Equation (3.2) gives

the form of the per observation expected Fisher information matrix as

$$I_i(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right].$$

The method for approximating the above information matrix uses a first-order Taylor series approximation of the second-order derivatives.

3.2.1 Method 3: Approximating the second-order derivatives

The (j,l)-th element of the per observation expected Fisher information matrix can be written as

$$I_i(\boldsymbol{\theta})_{(j,l)} \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial \theta_j \partial \theta_l}\right] \;=\; \mathbb{E}\left[f_{j,l}(a_i, \boldsymbol{x}_i)\right],$$

where

$$-\frac{\partial^2 \ln f_{Y_i}}{\partial \theta_j \partial \theta_l} \;=\; f_{j,l}(a_i, \boldsymbol{x}_i).$$

That is, the second-order derivatives are functions of a_i and $\boldsymbol{x}_i$. The information matrix for the i-th observation can then be approximated using the Taylor approximation derived in Section 3.1.2. The (j,l)-th element of the approximated per observation expected Fisher information matrix for the i-th observation is

$$\widehat{I}_i(\boldsymbol{\theta})_{(j,l)} \;=\; \mathbb{E}\left[\widehat{f}_{j,l}(a_i, \boldsymbol{x}_i)\right] \;=\; f_{j,l}\left(\mu_a, \boldsymbol{x}_i\right). \tag{3.6}$$

Thus the approximated information matrix can be calculated by evaluating any functions of a_i in the second-order derivatives at $a_i = \mu_a$ and multiplying the approximated derivative by minus one.

As with Section 3.1.2, a disadvantage of the form of this approximated per observation information matrix is that it is not guaranteed to be positive semidefinite. Hence estimation of the parameters, or subsets or linear combinations of the parameters may not be possible.

Calculating the approximated information matrix

The Taylor approximation of the information matrix in this section, given by equation (3.6), is just the Taylor approximation (3.5) given in Section 3.1.2. Therefore the comments regarding the calculation of the information matrix in Section 3.1.2 also apply here. The calculations differ in that, in Section 3.1.2, the approximations are applied to the product of the first-order derivatives and in this section, the approximations are applied to the second-order derivatives.

Chapter 4

Stochastic Frontier Models

In Chapter 2 the usual statistical model with one symmetrically distributed random error term was extended to a stochastic model consisting of two error terms, the usual random error term corresponding to statistical noise and an additional asymmetrically distributed error term. One particular example of this type of model can be found in the econometric literature and is called a 'stochastic frontier model.' The application of stochastic frontier models is in obtaining measures of efficiency that enable a comparison of performance across similar organisations. Inefficiency, a measure of the magnitude of sub-optimal performance, is represented by the asymmetric error term in a stochastic frontier model.

4.1 Measurement of Efficiency

Units producing outputs, such as goods or services, are commonly called producers, production units, decision making units or organisations. Because they are the units being observed, they are also referred to here as observational units. Production units can vary in size. For example, a production unit can be a staff member of a university, departments within a university, or universities

within a country.

A loose definition of efficiency is that efficiency is the relationship between what an organisation produces and what it could feasibly produce. Quantification of efficiency measures is useful for several reasons. Relative measures of efficiency facilitate comparisons across similar production units. Where inefficiency exists, further analysis can identify the factors causing inefficiency. Additionally, such an analysis informs policy decisions regarding improvement of efficiencies. It may be helpful to broadly distinguish between the different types of efficiency measures discussed in the literature.

4.1.1 Input-oriented versus output-oriented efficiency

Measures of efficiency can be input-oriented or output-oriented. When input quantities are fixed so that output varies across producers, the efficiency measure is *output-oriented* because the objective of producers is to maximise output. When output quantities are fixed so that inputs vary across producers, the efficiency measure is *input-oriented* because the objective of producers is to best allocate input quantities and minimise input usage.

4.1.2 Technical and economic efficiency

If the only information available are input and output quantities, that is, there is no information on input or output prices, then the type of efficiency that can be measured is *technical efficiency*. Technical efficiency can be input-oriented or output-oriented, with output-oriented technical efficiency being the more common measure of the two. Input-oriented technical inefficiency occurs when more resources than are required are used to produce a given amount of outputs. Output-oriented technical inefficiency occurs when the amount of outputs produced is less than the maximum amount possible for a given amount

of resources. Technical efficiency is also known as *X-efficiency*.

If price information on the inputs and outputs is available, in addition to input and output quantities, then *economic efficiency* can be measured. Economic efficiency is the more general term when some form of pricing information is also available. Specific types of economic efficiency include *cost efficiency*, *profit efficiency* and *revenue efficiency*. The type of economic efficiency that is measured will depend on the behavioural objective imposed on producers. For example, whether the objective of producers is to minimise costs or maximise profits or revenue. Economic efficiency measures are input-oriented. It is possible to decompose economic efficiency into technical efficiency and allocative efficiency. That is, if additional information is available on prices then it is possible to obtain a measure of *allocative efficiency* in addition to technical efficiency.

$$\textit{Economic Efficiency} \quad = \quad \textit{Technical Efficiency} + \textit{Allocative Efficiency}$$

Allocative inefficiency is input-oriented and occurs when the mixture of inputs used is not the mixture with the lowest possible cost for producing a given amount of outputs.

4.1.3 Frontiers and relative efficiency

It is important to note that efficiency is a relative measure. Vast amounts of literature pertaining to the measurement of efficiency are based on the concept of a 'frontier'. The development of frontier models began with Koopmans's (1951) and Debreu's (1951) definitions of efficiency. Influenced by these definitions, Farrell (1957) was the first to measure efficiency empirically and propose a decomposition of economic efficiency into technical efficiency and allocative efficiency. The relative efficiency of a producer can be measured relative to a frontier and hence relative to other producers. There are different types of frontiers corresponding to the different types of efficiency measures discussed above.

A *production frontier* is a graph of the maximum feasible output producible given fixed resources. Hence a production frontier envelopes producer outputs from above. If what a producer actually produces is less than what it could feasibly produce then it will lie below the frontier. The further below the production frontier a producer lies, the more inefficient it is. The type of efficiency that can be measured using a production frontier is *technical efficiency*.

A *cost frontier* is a graph of the minimum feasible cost for producing a fixed amount of outputs. Hence a cost frontier envelopes producer costs from below. The further above the cost frontier a producer lies, the more inefficient they are. The type of efficiency that can be measured using a cost frontier is *cost efficiency*.

Other types of frontiers include *profit frontiers* and *revenue frontiers*. Profit efficiency is measured relative to a profit frontier and revenue efficiency is measured relative to a revenue frontier.

4.1.4 Parametric versus nonparametric efficiency analysis

If the frontier has a functional form, that is, if a parametric model for the frontier can be formulated, then several parametric approaches have been developed in the literature for obtaining measurements of efficiency. The type of parametric technique employed will depend on whether the frontier model is *deterministic* (no random error in the model) or *stochastic* (random error in the model). However it has been clearly established that stochastic frontier models are superior to deterministic frontier models (Aigner, Lovell & Schmidt 1977). Parametric analyses of deterministic frontier models include *goal programming*, modified versions of *ordinary least squares estimation* and *maximum likelihood estimation* (e.g. Aigner et al. 1977, Fried, Lovell & Schmidt 1993, Kumbhakar & Lovell 2000). For stochastic frontier models, the parametric method of analysis is called *Stochastic Frontier Analysis* (SFA) (e.g. Fried et al. 1993, Kumbhakar & Lovell 2000, Jacobs, Smith & Street 2006).

If a suitable parametric model for the frontier cannot be specified then non-parametric approaches for obtaining efficiency measurements are readily available. The most popular nonparametric technique is called *Data Envelopment Analysis* (DEA) (e.g. Charnes, Cooper, Lewin & Seiford 1994, Cooper, Seiford & Tone 2000, Cooper, Seiford & Zhu 2004).

The primary focus of this chapter is on the stochastic frontier analysis of single-output cross-sectional stochastic production frontier models used to obtain measures of output-oriented technical efficiency. The alternative methods listed above shall be discussed briefly first to provide a historical perspective on the development of frontier models.

4.2 Deterministic Production Frontier Models and Technical Efficiency

The model considered in this section is restricted to a single-output production frontier for cross-sectional data. For the i-th observational unit, the production frontier model is

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot TE_i, \qquad i = 1, \ldots, N, \tag{4.1}$$

where the observed response y_i is a scalar output, $\boldsymbol{x}_i$ is a vector of m inputs, $\boldsymbol{\beta}$ is a vector of p unknown technology parameters, $f(\boldsymbol{x}_i, \boldsymbol{\beta})$ is the deterministic production frontier and TE_i is the output-oriented technical efficiency. For a first-order model $p = m + 1$.

Technical efficiency of the i-th observational unit is the ratio of observed output to maximum feasible output

$$TE_i = \frac{y_i}{f(\boldsymbol{x}_i, \boldsymbol{\beta})}. \tag{4.2}$$

If the observed output y_i reaches its maximum obtainable value $f(\boldsymbol{x}_i, \boldsymbol{\beta})$ then $TE_i = 1$. That is, the producer is operating at the frontier of production and is 100% efficient. Values of $TE_i < 1$ measure the shortfall of observed output from maximum feasible output. Note that model (4.1) is *deterministic* (contains no statistical noise). Therefore, from equation (4.2), any shortfall in output y_i from maximum feasible output $f(\boldsymbol{x}_i, \boldsymbol{\beta})$ is solely attributable to the inefficiency of the producer. Letting

$$TE_i \;=\; \exp\{-u_i\}, \qquad u_i \geq 0,$$

will ensure that $0 \leq TE_i \leq 1$ and that observed output y_i for the i-th producer will lie below the frontier $f(\boldsymbol{x}_i, \boldsymbol{\beta})$, that is

$$y_i \;\leq\; f(\boldsymbol{x}_i, \boldsymbol{\beta}).$$

Equation (4.1) can then be rewritten as

$$y_i \;=\; f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{-u_i\}, \qquad u_i \geq 0,$$

where u_i represents the shortfall of output from the frontier for each observational unit. If productive technology takes a log-linear Cobb-Douglas form (Cobb & Douglas 1928) then the single-output deterministic production frontier model can be represented as

$$\ln y_i \;=\; \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_{ij} - u_i. \tag{4.3}$$

Deterministic techniques, such as goal programming, can be applied to model (4.3) to *calculate* the parameter vector $\boldsymbol{\beta}$, the vector u_i and hence the technical efficiency TE_i. If a distributional assumption is placed on u_i, deterministic econometric techniques, such as corrected ordinary least squares (COLS), modified ordinary least squares (MOLS) and maximum likelihood estimation (MLE), can be applied to *estimate* the parameter vector and obtain estimates of u_i and hence of TE_i.

4.2.1 Goal programming

Aigner & Chu (1968) were the first to calculate the parameter vector $\boldsymbol{\beta}$ in model (4.3) by solving deterministic optimisation problems. Model (4.3) can be converted to either a linear programming (LP) model or a quadratic programming (QP) model. The parameters are calculated using mathematical programming techniques, rather than estimated in any statistical sense. Hence the drawback of this method is that it precludes any statistical inference concerning the calculated parameters.

4.2.2 Maximum Likelihood Estimation (MLE)

If a distributional assumption is imposed on the u_i, maximum likelihood estimates of the parameters in model (4.3) can be obtained along with a measure of their precision. Schmidt (1976) showed that if the u_i are exponentially distributed, the maximum likelihood estimates of the parameters are the parameter values calculated using the linear programming model. If the u_i are half normally distributed, the maximum likelihood estimates of the parameters are the parameter values calculated using the quadratic programming model.

Greene (1980a) showed that the Hessians of the log-likelihood functions are singular under the exponential and half normal distributions for the deterministic production frontier and proposed an alternative model where u_i is gamma distributed. However, there is no equivalent mathematical programming problem for a gamma distributed deterministic frontier model.

4.2.3 Corrected Ordinary Least Squares (COLS)

Winsten's (1957) discussion on Farrell's (1957) paper suggests a two step approach to estimate the deterministic production frontier. Step one is to obtain the ordinary least squares (OLS) estimates. The OLS estimates $\hat{\beta}_i$ $(i \neq 0)$ of the

slope parameters are consistent and unbiased but the estimate $\hat{\beta}_0$ for the intercept, although consistent, is biased. Under OLS there will be some producers who will lie above the frontier. This is undesirable as it implies that some producers are outputting more than the maximum that is feasible. Step two 'corrects' this by shifting up the intercept so that the estimated frontier bounds the data from above. The COLS estimate of the intercept is

$$\hat{\beta}_0^* = \hat{\beta}_0 + \max_i\{\hat{u}_i\},$$

where the $\hat{u}_i$ are the OLS residuals. Using this correction, at least one producer will lie on the frontier with the remaining producers lying below the frontier. The OLS residuals also require correction giving the COLS residuals

$$\hat{u}_i^* = \max_i\{\hat{u}_i\} - \hat{u}_i.$$

Technical efficiency is then estimated using

$$TE_i = \exp\{-\hat{u}_i^*\}.$$

Because only the OLS intercept is corrected, the estimated COLS frontier is parallel to the fitted OLS regression line. Applying the same correction to all producers implies that the structure of 'best practice' production technology is the same as the structure of 'central tendency' production technology. This restrictive property of COLS is undesirable as the production technology of best practice producers should be permitted to differ from the production technology of less efficient producers.

4.2.4 Modified Ordinary Least Squares (MOLS)

A variation of the COLS procedure proposed by Afriat (1972) and Richmond (1974) assumes that the $u_i > 0$ follow an asymmetric distribution such as the half normal or exponential. As with COLS, the first step is to estimate the parameters

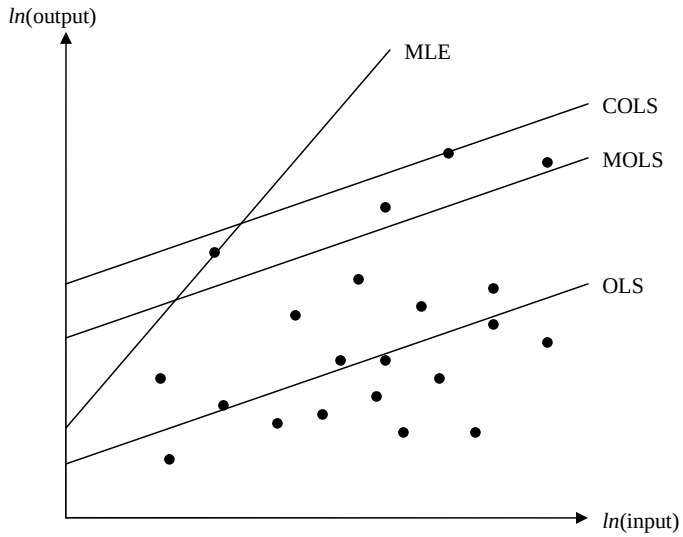

Figure 4.1: MLE, COLS and MOLS deterministic production frontiers.

via OLS. The 'modification' applied in the second step shifts up the intercept by the mean of the assumed one-sided distribution. The respective MOLS estimates of the intercept and the residuals are

$$\begin{aligned} \hat{\beta}_0^{**} &= \hat{\beta}_0 + \mathbb{E}[\hat{u}_i], \\ \hat{u}_i^{**} &= \mathbb{E}[\hat{u}_i] - \hat{u}_i. \end{aligned}$$

Technical efficiency is then estimated using

$$TE_i = \exp\{-\hat{u}_i^{**}\}.$$

Shifting up the intercept by $\mathbb{E}[\hat{u}_i]$ will not guarantee that the frontier is shifted up far enough to bound all producers from above. If a producer has a sufficiently large positive OLS residual it is possible that $(\mathbb{E}[\hat{u}_i] - \hat{u}_i) < 0$. Conversely, it is possible that the frontier may be shifted up too far so that no producer is close to the frontier, hence no producer is technically efficient. Additionally, as in the COLS case, the MOLS frontier is parallel to the fitted OLS regression line.

MLE, OLS, COLS and MOLS are illustrated in Figure 4.1. An ordinary least

squares approach does not allow for technical inefficiency because any variation in outputs not associated with variation in inputs is due to statistical noise. Conversely, all the above techniques which are applied to the deterministic production frontier model (4.3) attribute any shortfall in output entirely to factors within the control of the producer. A deterministic model does not permit the amount of output to vary due to random events that are out of the control of the producer. Clearly what is required is a model that attributes variation in outputs not due to variation in inputs to a combination of of both inefficiency (controllable by a producer) and statistical noise (random events outside the control of the producer). A stochastic production frontier model is one such model.

4.3 Stochastic Production Frontier Models and Technical Efficiency

As with the deterministic model, only cross-sectional data, which are observed at a single point in time, shall be considered in detail. Panel (or longitudinal) data, which are taken over several time points, shall be discussed briefly later. Also, the stochastic frontier under consideration is restricted to a single-output production frontier.

Meeusen & van den Broeck (1977) and Aigner et al. (1977) independently developed a stochastic production frontier model which improved on the deterministic frontier model of Aigner & Chu (1968) by allowing random events to contribute to variations in producer output. For the i-th observational unit, the stochastic frontier model is

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\} \cdot TE_i, \qquad i = 1, \ldots, N, \tag{4.4}$$

where the observed response y_i is a scalar output, $\boldsymbol{x}_i$ is a vector of m inputs, $\boldsymbol{\beta}$ is a vector of p unknown technology parameters, $f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\}$ is the stochastic

production frontier and TE_i is the output-oriented technical efficiency. Stochasticity of the frontier is due to the term v_i which represents statistical noise and is intended to capture the effects of random events beyond the control of the producer. Hence the v_i are assumed to be identically and independently distributed with mean zero. Technical efficiency of the i-th observational unit is the ratio of observed output to maximum feasible output

$$TE_i = \frac{y_i}{f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\}}. \tag{4.5}$$

If the observed output y_i reaches its maximum obtainable value $f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\}$, accounting for random error, then $TE_i = 1$ and the producer is 100% efficient. By rearranging equation (4.5) to

$$TE_i \cdot \exp\{v_i\} = \frac{y_i}{f(\boldsymbol{x}_i, \boldsymbol{\beta})},$$

and comparing with equation (4.2), it is clear that the advantage of the stochastic specification of the frontier is that it allows any shortfall in realised output to be attributable to both technical inefficiency and random events experienced by the producer. Letting technical efficiency take the same form as in the deterministic model

$$TE_i = \exp\{-u_i\}, \qquad u_i \geq 0,$$

will ensure that $0 \leq TE_i \leq 1$ and that the observed output y_i for the i-th producer will lie below the stochastic frontier $f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\}$, that is

$$y_i \leq f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\}.$$

Equation (4.4) can be rewritten as

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) \cdot \exp\{v_i\} \cdot \exp\{-u_i\}, \qquad u_i \geq 0,$$

where u_i represents the difference between the realised output and maximum output for each observational unit. That is, u_i represents technical inefficiency

and is assumed to have an asymmetric distribution. The most frequently used distributions for u_i are the half normal and exponential distributions truncated from below at zero. Assuming productive technology takes a log-linear Cobb-Douglas form, the single-output stochastic frontier model can be represented as

$$\ln y_i = \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_{ij} + v_i - u_i.$$

The overall error $\varepsilon_i = v_i - u_i$ is often referred to as a 'composed error' term, composed of the traditional symmetric random noise component v_i and an additional one-sided inefficiency component u_i. The two error terms v_i and u_i are assumed to be independent of each other and of the input variables. Separation of the two error terms allows for efficiency measurement analysis.

Estimation by OLS provides consistent estimates $\hat{\beta}_i$ $(i \neq 0)$ of the slope parameters but the estimate $\hat{\beta}_0$ of the intercept is inconsistent. Additionally, OLS considers only the composed error ε_i and hence does not provide estimates of technical efficiency for each producer. OLS is useful however in providing a test for the presence of inefficiency. If $\varepsilon_i = v_i$ then $u_i = 0$ and the OLS residuals will be symmetrically distributed suggesting an absence of technical inefficiency in the data. Schmidt & Lin (1984) and Coelli (1995) provide alternative test statistics to test for possible inefficiency in the data by testing if the data are skewed using the second and third sample moments of the OLS residuals. Coelli's (1995) test statistic is more commonly used as it is asymptotically distributed as $N(0, 1)$.

While OLS can be used to provide consistent estimates of the slope parameters, additional assumptions and a different estimation technique are required to obtain consistent estimates of the intercept and estimates of technical efficiency for each producer. Estimation of all the β parameters and the u_i can however be achieved via maximum likelihood.

Let random variables U and V have respective realisations u and v, where u is

associated with technical efficiency and v is associated with statistical noise. Estimation of the parameters can be achieved under maximum likelihood estimation if we assume that random variables U and V are distributed as follows

(i) $U \sim$ asymmetric i.i.d, e.g. nonnegative half normal, exponential, nonnegative truncated normal, gamma

(ii) $V \sim N(0, \sigma_v^2)$ i.i.d.

(iii) U and V are distributed independently of each other, and of the input variables.

Assumption (i), for the half normal and exponential distributions, is based on the premise that the modal value of technical inefficiency is zero, with increasing values of technical inefficiency becoming increasingly less likely. The truncated normal and gamma specifications allow a nonzero modal value of technical inefficiency, but still with the premise that increasing values of inefficiency are increasingly less likely. Assumption (ii) assumes that random error is normally distributed with zero mean and constant variance. The second part of assumption (iii) can be problematic since if producers have knowledge of their technical efficiency, this may influence their choice of inputs to production. This assumption is relaxed when measurements are taken over time.

The log-likelihoods, their derivatives and information matrices were derived in Chapter 2 for the more general model

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + c_u u_i + c_v v_i, \qquad \{c_u, c_v\} \in \mathbb{R}.$$

Applying a logarithmic transformation to the response y_i and predictors $\boldsymbol{x}_i$, and letting

$$f(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}, \tag{4.6}$$

$$c_u = -1, \tag{4.7}$$

$$c_v = 1, \tag{4.8}$$

gives the log-linear Cobb-Douglas form of the single-output stochastic production frontier model

$$\begin{aligned} \ln y_i &= \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_{ij} + v_i - u_i \\ &= \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta} + v_i - u_i. \end{aligned} \tag{4.9}$$

In the above equation $\boldsymbol{f}^T(\boldsymbol{x}_i) = (1, \ln x_{i1}, \ln x_{i2}, \ldots, \ln x_{im})$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ and the overall composed error is $\varepsilon_i = v_i - u_i$. Hence the maximum likelihood estimates and information matrices can be derived for the stochastic production frontier model by substituting equations (4.6) to (4.8) into the equations derived in Chapter 2. In the remainder of this chapter, y_i will be the notation used for the response (output) although when applying a log-linear Cobb-Douglas stochastic frontier model to data, logarithms must be applied to the inputs and outputs.

4.3.1 Normal-half normal model

Let U follow a nonnegative half normal distribution, that is $U \sim N^+(0, \sigma_u^2)$. Properties of the more general form of the composed error term where $E = c_u U + c_v V$ were derived in Section 2.2 of Chapter 2. The probability density functions $f_U(u)$ and $f_V(v)$ are given in equations (2.3) and (2.4) respectively with their joint density $f_{U,V}(u, v)$ given in equation (2.6).

For the stochastic production frontier model where the composed error takes the form $E = V - U$, the joint density (2.7) simplifies to

$$f_{U,E}(u, \varepsilon) = \frac{1}{\pi \sigma_u \sigma_v} \exp\left\{ -\frac{u^2}{2\sigma_u^2} - \frac{(\varepsilon + u)^2}{2\sigma_v^2} \right\}.$$

The marginal density (2.8) simplifies to

$$f_E(\varepsilon) = \frac{2}{\sigma_G} \phi\left(\frac{\varepsilon}{\sigma_G}\right) \Phi\left(-\frac{\lambda \varepsilon}{\sigma_G}\right),$$

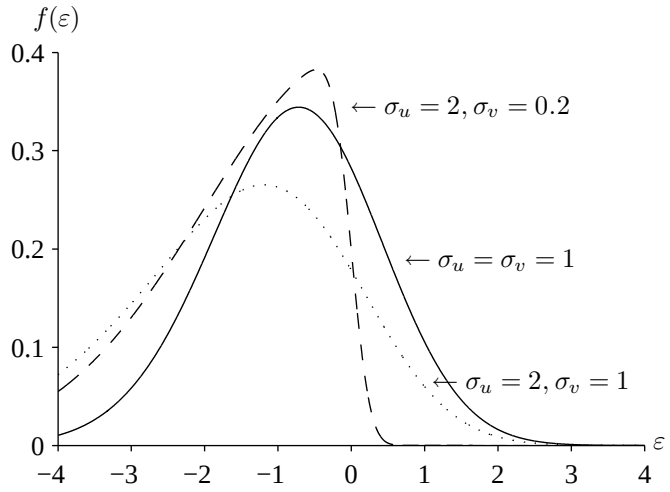

Figure 4.2: Normal-half normal distributions.

with mean and variance from equations (2.9) and (2.10) simplifying to give

$$\mathbb{E}[E] = -\sqrt{\frac{2}{\pi}}\sigma_u,$$

$$Var(E) = \frac{\pi - 2}{\pi}\sigma_u^2 + \sigma_v^2,$$

where $\sigma_G^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda = \sigma_u/\sigma_v$. Three different normal-half normal distributions are plotted in Figure 4.2. All distributions are negatively skewed with negative modes and means since $\sigma_u > 0$ for each density. The reparameterisation of σ_u and σ_v to λ gives an indication of the relative contribution of u and v to ε. As $\lambda \to 0$ either $\sigma_v^2 \to \infty$ or $\sigma_u^2 \to 0$, so that statistical noise dominates the term associated with technical efficiency. As $\lambda \to \infty$ either $\sigma_u^2 \to \infty$ or $\sigma_v^2 \to 0$, so that the technical efficiency component dominates the statistical noise in the determination of ε. Coelli (1995) gives the appropriate one-sided likelihood ratio test statistic for testing the hypothesis that $\lambda = 0$.

To calculate an estimate of mean technical efficiency for all producers, or estimates of the technical efficiency for each individual producer, the parameters must first be estimated. The log-likelihood function (2.11) for N observational

units simplifies to

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \left\{ \ln\left(\frac{2}{\sigma_G}\right) + \ln \phi\left(\frac{y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}}{\sigma_G}\right) + \ln \Phi(-a_i) \right\},$$

where

$$a_i = \frac{\lambda \varepsilon_i}{\sigma_G} = \frac{\lambda[y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}]}{\sigma_G}.$$

Maximum likelihood estimates of the parameters can be obtained by maximising the log-likelihood function with respect to the parameters.

The mean technical efficiency of all producers is given by

$$\mathbb{E}[\exp\{-U\}] = 2[1 - \Phi(\sigma_u)] \exp\left\{\frac{\sigma_u^2}{2}\right\},$$

(Lee & Tyler 1978). This estimator is preferred to Aigner et al.'s (1977) original estimator $(1 - \mathbb{E}[U])$ which is only the first-order term in the Taylor series expansion of $\exp\{-U\}$. The average technical efficiency of all producers is not usually of primary interest. Estimates of individual producer efficiencies are desirable to enable comparison across producers.

From equation (C.12) in Appendix C, the conditional density of U given E can be rewritten as

$$f_{U|E}(u|\varepsilon) = \frac{\frac{1}{\sigma_*}\phi\left(\frac{u - \mu_*}{\sigma_*}\right)}{\Phi\left(\frac{\mu_*}{\sigma_*}\right)}, \tag{4.10}$$

where $\mu_* = -\frac{\sigma_u^2}{\sigma_G^2}\varepsilon$ and $\sigma_* = \frac{\sigma_u \sigma_v}{\sigma_G}$. The conditional density $f_{U|E}(u|\varepsilon)$ is distributed as $N^+(\mu_*, \sigma_*^2)$, hence the mean or the mode of this distribution can be used as a point estimator for u_i. Equations (C.13) and (C.14) simplify to give the conditional mean and mode for the i-th observational unit as

$$\mathbb{E}[U_i|E_i] = \mu_{*i} + \frac{\sigma_* \phi\left(-\frac{\mu_{*i}}{\sigma_*}\right)}{\Phi\left(\frac{\mu_{*i}}{\sigma_*}\right)} = \sigma_* \left[-a_i + \frac{\phi(a_i)}{\Phi(-a_i)}\right], \tag{4.11}$$

$$M(U_i|E_i) \;=\; \begin{cases} \mu_{*i} & \text{if } \mu_{*i} \geq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

Estimates of technical efficiency for each observational unit can then be obtained from

$$TE_i \;=\; \exp\{-\hat{u}_i\}, \tag{4.13}$$

where $\hat{u}_i$ is $\mathbb{E}[U_i|E_i]$ or $M(U_i|E_i)$. The point estimator $\hat{u}_i = \mathbb{E}[U_i|E_i]$ was originally proposed by Jondrow, Lovell, Materov & Schmidt (1982). Battese & Coelli (1988) provide an alternative estimator

$$TE_i \;=\; \mathbb{E}[\exp\{-U_i\}|E_i] \;=\; \left[\frac{\Phi\left(-\sigma_* + \dfrac{\mu_{*i}}{\sigma_*}\right)}{\Phi\left(\dfrac{\mu_{*i}}{\sigma_*}\right)}\right] \exp\left\{-\mu_{*i} + \tfrac{1}{2}\sigma_*^2\right\}. \tag{4.14}$$

Jondrow et al.'s estimator is the first-order term in the Taylor series expansion of Battese & Coelli's estimator, hence Battese & Coelli's estimator is usually preferred. Horrace & Schmidt (1996), Bera & Sharma (1999) and Hjalmarsson, Kumbhakar & Heshmati (1996) obtained confidence intervals for Jondrow et al.'s estimator. Bera & Sharma also obtained confidence intervals for Battese & Coelli's estimator.

The information matrix for the log-linear Cobb-Douglas stochastic production frontier model can be derived by substituting $c_u = -1$ and $c_v = 1$ into the corresponding equations in Section 2.2 of Chapter 2. The information matrix is required to obtain standard errors for the maximum likelihood parameter estimates. It can also be used to design experiments for the frontier model. The information matrix and the derivatives used to calculate it are given in Appendix B.1.

4.3.2 Normal-exponential model

Let U follow an exponential distribution with scale parameter σ_u (the inverse scale $1/\sigma_u$ is called the rate parameter), that is $U \sim \text{Exponential}(1/\sigma_u)$. Properties of the more general form of the composed error term where $E = c_u U + c_v V$ were derived in Section 2.3 of Chapter 2. The probability density functions $f_U(u)$ and $f_V(v)$ are given in equations (A.1) and (2.4) respectively with their joint density $f_{U,V}(u, v)$ given in equation (A.2).

For the stochastic production frontier model where the composed error takes the form $E = V - U$, the joint density (A.3) simplifies to

$$f_{U,E}(u, \varepsilon) = \frac{1}{\sqrt{2\pi}\sigma_u\sigma_v} \exp\left\{-\frac{u}{\sigma_u} - \frac{(\varepsilon + u)^2}{2\sigma_v^2}\right\}.$$

The marginal density (2.14) simplifies to

$$f_E(\varepsilon) = \frac{1}{\sigma_u} \exp\left\{\frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2}\right\} \Phi\left(-\frac{\varepsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_u}\right),$$

with mean and variance from equations (2.15) and (2.16) simplifying to give

$$\begin{aligned} \mathbb{E}[E] &= -\sigma_u, \\ Var(E) &= \sigma_u^2 + \sigma_v^2. \end{aligned}$$

Three different normal-exponential distributions are plotted in Figure 4.3. The shape of the distribution will depend on σ_u and σ_v. As the ratio $\sigma_u/\sigma_v \to \infty$ the density looks increasingly more like a negative exponential distribution. As $\sigma_u/\sigma_v \to 0$ the density looks increasingly more like a normal distribution.

The parameters can be estimated by maximising the log-likelihood function. The log-likelihood function (2.17) simplifies to

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{1=1}^{N} \left\{\ln\left(\frac{1}{\sigma_u}\right) + \frac{y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} + \ln \Phi(-a_i)\right\},$$

where

$$a_i = \frac{\varepsilon_i}{\sigma_v} + \frac{\sigma_v}{\sigma_u} = \frac{y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}}{\sigma_v} + \frac{\sigma_v}{\sigma_u}.$$

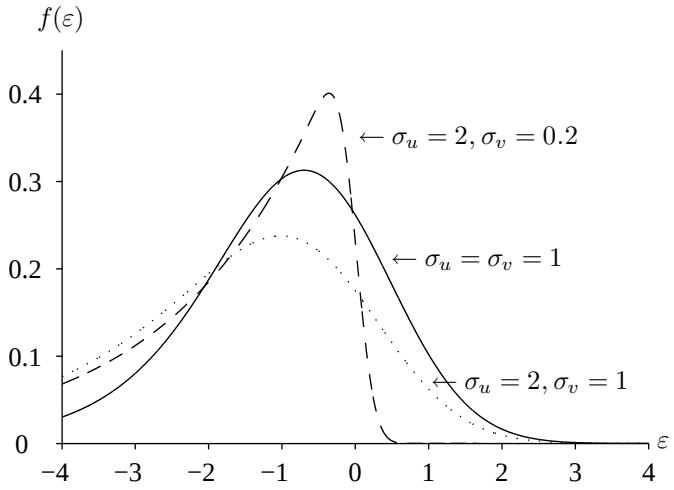

Figure 4.3: Normal-exponential distributions.

Given the maximum likelihood parameter estimates, producer specific estimates of technical efficiency can be obtained from the conditional density of U given E, which is given in equation (4.10), with

$$\mu_* = -\varepsilon - \frac{\sigma_v^2}{\sigma_u} \qquad \text{and} \qquad \sigma_* = \sigma_v.$$

The conditional mean and mode for the i-th observational unit can be calculated using equations (4.11) and (4.12) respectively. Estimates of technical efficiency for each observational unit can then be obtained using the conditional mean or mode to estimate u_i and substituting this into equation (4.13), or by using equation (4.14). Confidence intervals can also be derived in the same manner as the normal-half normal case. The derivatives of the log-likelihood function are given in Appendix B.2 along with the information matrix.

4.3.3 Normal-truncated normal model

Let U follow a nonnegative truncated normal distribution, that is $U \sim N^+(\mu, \sigma_u^2)$ where μ is the mode. If the distribution was not truncated, μ would be both the mean and the mode. When $\mu = 0$ the nonnegative truncated normal

distribution simplifies to the nonnegative half normal distribution. Hence the nonnegative truncated normal distribution generalises the nonnegative half normal distribution by allowing the modal value of technical efficiency to be nonzero, thus permitting a more flexible structure for the pattern of efficiency in the data. The normal-truncated normal formulation for a stochastic production frontier was introduced by Stevenson (1980).

Properties of the more general form of the composed error term where $E = c_u U + c_v V$ were derived in Section 2.4 of Chapter 2. The probability density functions $f_U(u)$ and $f_V(v)$ are given in equations (A.4) and (2.4) respectively with their joint density $f_{U,V}(u,v)$ given in equation (A.5). For the stochastic production frontier model where the composed error takes the form $E = V - U$, the joint density (A.6) simplifies to

$$f_{U,E}(u,\varepsilon) = \frac{1}{2\pi\sigma_u\sigma_v}\exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2}-\frac{(\varepsilon+u)^2}{2\sigma_v^2}\right\}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}.$$

The marginal density (2.22) simplifies to

$$f_E(\varepsilon) = \frac{1}{\sigma_G}\phi\left(\frac{\mu+\varepsilon}{\sigma_G}\right)\Phi\left(\frac{\mu}{\lambda\sigma_G}-\frac{\lambda\varepsilon}{\sigma_G}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1},$$

with mean and variance from equations (2.23) and (2.24) simplifying to give

$$\mathbb{E}[E] = -\mu - h\left(-\frac{\mu}{\sigma_u}\right)\sigma_u,$$

$$Var(E) = \left\{1-\frac{\mu}{\sigma_u}h\left(-\frac{\mu}{\sigma_u}\right)-\left[h\left(-\frac{\mu}{\sigma_u}\right)\right]^2\right\}\sigma_u^2+\sigma_v^2,$$

where $\sigma_G^2 = \sigma_u^2 + \sigma_v^2$ and $\lambda = \sigma_u/\sigma_v$. Three different normal-truncated normal distributions are plotted in Figure 4.4 where $\sigma_u = \sigma_v = 1$ for all densities and μ is negative, zero (the normal-half normal case) and positive.

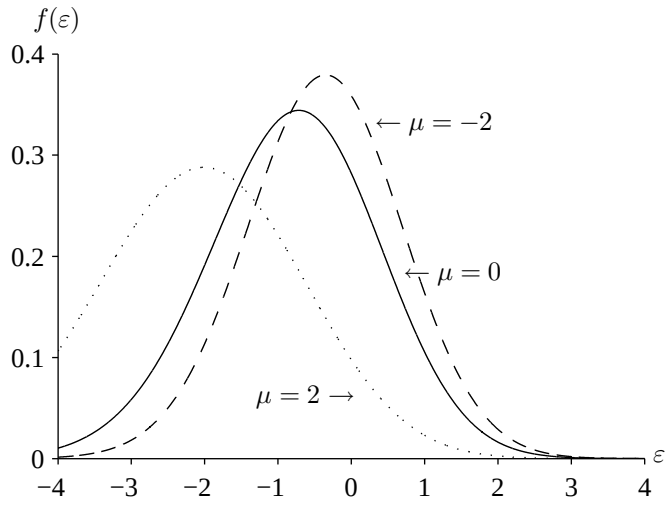

Figure 4.4: Normal-truncated normal distributions with $\sigma_u = \sigma_v = 1$.

The parameters can be estimated by maximising the log-likelihood function. The log-likelihood function (2.25) simplifies to

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \left\{ -\ln \sigma_G + \ln \phi \left(\frac{\mu + [y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}]}{\sigma_G} \right) + \ln \Phi(-a_{1i}) - \ln \Phi\left(-a_2\right) \right\},$$

where

$$a_{1i} = -\frac{\mu}{\lambda \sigma_G} + \frac{\lambda \varepsilon_i}{\sigma_G} = -\frac{\mu}{\lambda \sigma_G} + \frac{\lambda [y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}]}{\sigma_G},$$

$$a_2 = -\frac{\mu}{\sigma_u} = -\frac{\mu(\lambda^2 + 1)^{1/2}}{\lambda \sigma_G},$$

$$\sigma_u = \frac{\lambda \sigma_G}{(c_u^2 \lambda^2 + c_v^2)^{1/2}}.$$

Once the maximum likelihood estimates have been calculated, producer specific estimates of technical efficiency can be obtained from the conditional density of U given E, which is given in equation (4.10), with

$$\mu_* = \frac{\sigma_v^2 \mu - \sigma_u^2 \varepsilon}{\sigma_G^2} \quad \text{and} \quad \sigma_* = \frac{\sigma_u \sigma_v}{\sigma_G}.$$

Estimates of efficiency can then be calculated in the same manner as for the normal-half normal and normal-exponential cases. The calculations for the information matrix are given in Appendix B.3.

4.3.4 Normal-gamma model

Let U follow a gamma distribution with shape parameter α and scale parameter σ_u (the inverse scale $1/\sigma_u$ is called the rate parameter), that is $U \sim \text{Gamma}(\alpha, \sigma_u)$. When $\alpha = 1$ the gamma distribution simplifies to the exponential distribution. Hence the gamma distribution generalises the exponential distribution by allowing the shape parameter to take a value other than one. Different values of the shape parameter will produce densities with different modal values of technical efficiency (the modal value is zero for the exponential distribution) thus permitting a more flexible structure for the pattern of efficiency in the data. The normal-gamma formulation was introduced by Greene (1980*a*), Greene (1980*b*) and Stevenson (1980), and later extended by Greene (1990).

Properties of the more general form of the composed error term where $E = c_u U + c_v V$ were derived in Section 2.5 of Chapter 2. The probability density functions $f_U(u)$ and $f_V(v)$ are given in equations (A.7) and (2.4) respectively with their joint density $f_{U,V}(u, v)$ given in equation (A.8). For the stochastic production frontier model where the composed error takes the form $E = V - U$, the joint density (A.9) simplifies to

$$f_{U,E}(u, \varepsilon) = \frac{u^{\alpha-1}}{\Gamma(\alpha)\sigma_u^{\alpha}\sqrt{2\pi}\sigma_v} \exp\left\{-\frac{u}{\sigma_u} - \frac{(\varepsilon+u)^2}{2\sigma_v^2}\right\}.$$

The marginal density (2.28) simplifies to

$$f_E(\varepsilon) = \frac{1}{\Gamma(\alpha)\sigma_u^{\alpha}} \exp\left\{\frac{\varepsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2}\right\} \Phi\left(-\frac{\varepsilon}{\sigma_v} - \frac{\sigma_v}{\sigma_u}\right) \mathbb{E}[Q^{\alpha-1}],$$

where $Q \sim N^+\left(-\varepsilon - \sigma_v^2/\sigma_u, \sigma_v^2\right)$ and $\mathbb{E}[Q^{\alpha-1}]$ is a fractional moment of the non-negative truncated normal distribution of Q. The integration inherent in the last

two terms of $f_E(\varepsilon)$ poses some problems in estimation. Numerical approximation is required to evaluate the integral and the estimates can be sensitive to the quadrature rule used (Greene 1990). Stevenson (1980) and Beckers & Hammond (1987) provide two alternative (but equivalent) representations of the marginal density of E. Stevenson also gives a closed form expression for the normal-gamma density for $\alpha = 2$ and $\alpha = 3$. However, integer values of α restrict the gamma distribution to the Erlang distribution. Beckers & Hammond's formulation does not restrict α to integer values.

The mean and variance from equations (2.29) and (2.30) simplify to give

$$\begin{aligned} \mathbb{E}[E] &= -\alpha\sigma_u, \\ Var(E) &= \alpha\sigma_u^2 + \sigma_v^2. \end{aligned}$$

Three different normal-gamma distributions are plotted in Figure 4.5 where $\sigma_u = \sigma_v = 1$ for all densities and $\alpha = 1, 2, 3$. These values of α are convenient for illustrative purposes; when $\alpha = 1$, $\mathbb{E}[Q^{\alpha-1}] = 1$ and the density collapses to the normal-exponential density; when $\alpha = 2$, $\mathbb{E}[Q^{\alpha-1}]$ is the mean of the nonnegative truncated normal random variable Q; when $\alpha = 3$, $\mathbb{E}[Q^{\alpha-1}]$ can be obtained using the identity $\mathbb{E}[Q^2] = Var(Q) + \mathbb{E}[Q]^2$.

The parameters can be estimated by maximising the log-likelihood function. The log-likelihood function (2.31) simplifies to

$$\begin{aligned} \ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \Bigg\{ &- \ln \Gamma(\alpha) + \alpha \ln \left(\frac{1}{\sigma_u} \right) + \frac{\varepsilon_i}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \\ &+ \ln \Phi \left(-\frac{\varepsilon_i}{\sigma_v} - \frac{\sigma_v}{\sigma_u} \right) + \ln \mathbb{E}[Q_i^{\alpha-1}] \Bigg\}, \end{aligned}$$

where $\varepsilon_i = y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}$. Given the maximum likelihood parameter estimates, producer specific estimates of technical efficiency can be obtained from the conditional density of U given E. From equation (C.9) in Appendix C, the conditional

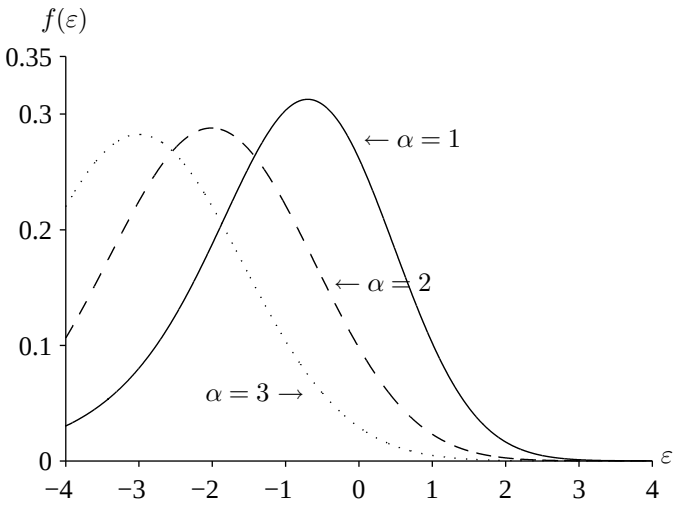

Figure 4.5: Normal-gamma distributions with $\sigma_u = \sigma_v = 1$.

density can be rewritten as

$$f_{U|E}(u|\varepsilon) = \frac{u^{\alpha-1}\dfrac{1}{\sigma_*}\phi\left(\dfrac{u-\mu_*}{\sigma_*}\right)}{\Phi\left(\dfrac{\mu_*}{\sigma_*}\right)\mathbb{E}[Q^{\alpha-1}]},$$

where $\mu_* = -\varepsilon - \dfrac{\sigma_v^2}{\sigma_u}$ and $\sigma_* = \sigma_v$. Equation (C.10) gives the conditional mean for the i-th observational unit as

$$\mathbb{E}[U_i|E_i] = \frac{\mathbb{E}[Q^{\alpha}]}{\mathbb{E}[Q^{\alpha-1}]},$$

which can be approximated numerically. Estimates of technical efficiency for each observational unit can then be obtained using $\hat{u}_i = \mathbb{E}[U_i|E_i]$ and substituting this into equation (4.13).

The information matrix and the derivatives used to calculate it are given in Appendix B.4. The elements of the expected information matrix involve complicated integrals which are inherent in the conditional expectations, variances and covariances appearing in the elements of the matrix. These quantities can be approximated numerically although the approximation error can pose a serious problem (Ritter & Simar 1997). Ritter & Simar (1997) also report that the

sample size needs to be in the hundreds to be able to estimate the shape parameter α which is required to obtain the efficiency estimates. Similar problems exist in estimating μ under the normal-truncated normal specification (Ritter & Simar 1997).

4.3.5 Sensitivity to distributional assumptions

Kumbhakar & Lovell (2000) investigate the concordance of correlation coefficients between efficiency rankings under the four distributional assumptions based on Greene's (1990) analysis of 123 U.S. electric utilities. Kumbhakar & Lovell report a strong rank correlation coefficient between the exponential and gamma estimates and also between the half normal and truncated normal estimates. This provides evidence to support Ritter & Simar's (1997) argument that the simpler half normal and exponential specifications should be implemented over the more flexible truncated normal and gamma distributions. Additionally, Kumbhakar & Lovell suggest that the efficiency estimates are generally not sensitive to the choice of one-parameter distribution (half normal or exponential).

Although the small number of empirical investigations which explore the sensitivity of *rankings* based on efficiency measurements report little sensitivity, they do not provide evidence on the sensitivity of the actual efficiency measurements themselves. It is only *suggested* that the actual efficiency measurements may *generally* be insensitive to the distributional assumptions. Nor do the studies discuss the sensitivity of the information matrix to the choice of distributional assumption imposed on the efficiency term.

4.3.6 Method of Moments Estimation

The details given above for the normal-half normal, normal-exponential, normal-truncated normal and normal-gamma specifications of composed error

were based on maximum likelihood estimation of the parameters. Like corrected ordinary least squares (COLS) and modified ordinary least squares (MOLS), maximum likelihood estimation (MLE) is carried out in two steps. In the first step, estimates of all the parameters are obtained via maximum likelihood. In the second step, estimates of technical efficiency are obtained conditional on the maximum likelihood estimates of the parameters by decomposing the maximum likelihood residuals into statistical noise and technical inefficiency.

An alternative estimation method is to obtain estimates of the model parameters using MOLS and then use equation (4.13) to obtain estimates of producer specific technical efficiency (Kumbhakar & Lovell 2000). Recall that the first step in MOLS estimation is to obtain consistent estimates of the slope parameters using ordinary least squares (OLS). In the second step of MOLS estimation, the second and third central moments of the OLS residuals can be used to estimate σ_u and σ_v. The estimate of σ_u is then used to obtain a consistent estimate of the intercept parameter. The estimated parameters are then used to obtain estimates of technical efficiency for each producer using equation (4.13). This procedure is referred to as 'method of moments estimation.'

Coelli's FRONTIER version 4.1 freeware for estimating stochastic frontier production and cost functions implements the method of moments approach described above. At the time of publication of this dissertation, FRONTIER version 4.1 was available for download free of charge with an accompanying manual at `http://www.uq.edu.au/economics/cepa/frontier.htm`. Sena (1999) reviews LIMPDEP 7.0 and FRONTIER 4.1 software used in the estimation of stochastic frontiers.

4.4 Extensions to Cross-Sectional Stochastic Production Frontier Models

The particular form of stochastic frontier model considered in detail above was restricted to a single-output production model for cross-sectional data.

4.4.1 Multiple-output stochastic distance functions

In situations where multiple inputs produce multiple outputs (rather than producing a single output), the single-output model can be extended to a multiple-output model using a *stochastic distance function* (Kumbhakar & Lovell 2000).

4.4.2 Stochastic production frontier models for panel data

(Schmidt & Sickles 1984) discuss three problems with cross-sectional stochastic production frontier models; (i) strong distributional assumptions are required for maximum likelihood estimation; (ii) maximum likelihood estimation requires that the technical inefficiency component u be independent of the regressors; and (iii) Jondrow et al.'s (1982) producer specific estimates of technical efficiency are not consistent. These limitations can be resolved if panel (or longitudinal) data are available (Kumbhakar & Lovell 2000). The stochastic production frontier model can be extended to allow data to be modelled over time with time-invariant technical efficiency (e.g. Pitt & Lee 1981, Schmidt & Sickles 1984, Kumbhakar 1987, Battese & Coelli 1988) or time-varying technical efficiency (e.g. Cornwell, Schmidt & Sickles 1990, Kumbhakar 1990, Lee & Schmidt 1993, Battese & Coelli 1992).

Linear mixed effects models with time-invariant technical efficiency

Assume we have N producers and that observations at times $t = 1, \ldots, T$ are collected for the i-th producer. Additionally, assume that there are no temporal trends. Cross-sectional model (4.9) can then be extended to a log-linear Cobb-Douglas stochastic production frontier model with time-invariant technical efficiency, which can be written as

$$\ln y_{it} \quad = \quad \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_{ijt} + v_{it} - u_i.$$

The above model is a linear mixed effects model. If the u_i are fixed then the model is a linear *fixed effects model* where the u_i are allowed to be correlated with the regressors or with the v_{it}. Since the u_i are fixed effects, they become producer specific intercept parameters.

When the u_i are randomly distributed with constant mean and variance, but are assumed to be uncorrelated with the regressors and with v_{it}, the above model is a linear *random effects model*, also called a *variance components model*. This one-way random effects model can be estimated by the standard generalised least squares method. If distributions on the u_i and v_{it} can be assumed, maximum likelihood estimation of the time-invariant model is possible and is structurally similar to the procedure applied to cross-sectional data. Note that the cross-sectional model, which is the primary focus of this thesis, can be viewed as a linear random effects model with a single observation collected at one time point for each producer.

As in the frontier literature (Coelli 1995), for mixed effects models, the variance ratio

$$\lambda^2 \quad = \quad \frac{\sigma_u^2}{\sigma_v^2},$$

sometimes referred to as the degree of correlation, is used in likelihood ratio testing of the variance components with null hypothesis $\lambda = 0$ (Stram & Lee

1994, Morrell 1998).

It is also worth noting that there is an existing literature on experimental design for variance component models. Khuri (2000) provides a comprehensive coverage of the literature on designs for estimating variance components. Mukerjee & Huda (1988) consider optimal design for the estimation of variance components, while Giovagnoli & Sebastiani (1989) discuss designs for estimation of both the mean and variance components. Aigner & Balestra (1988) present some work on optimum experimental design for error component models. Mentre, Mallet & Baccar (1997) report on optimal designs for estimating random effects regression models under cost constraints. Optimal Bayesian designs for one-way random effects models are explored in Lohr (1995). In more recent years Atkinson (2008) constructs optimum designs for random effects nonlinear regression models. However, the models dealt with in the design literature consider random effects with zero mean, whereas, stochastic production frontier models have a random effect u_i with nonzero mean.

4.4.3 Heteroskedasticity

It is not uncommon for the variance of the composed error term to be positively correlated with size-related characteristics of the observations, implying heteroskedasticity in the data. Heteroskedasticity can appear in either error component and can affect inferences about the model parameters and hence affect inferences about technical inefficiency. Kumbhakar & Lovell (2000) report that, for cross-sectional models: (i) unmodelled heteroskedasticity in v leads to biased estimates of technical efficiency although estimates of the model parameters are unbiased; (ii) unmodelled heteroskedasticity in u causes bias in both efficiency and model parameter estimates; and (iii) unmodelled heteroskedasticity in both error components causes bias in opposite directions, so there is hope that the overall bias may be small.

4.4.4 New developments: Bayesian techniques

Use of Bayesian techniques provides the researcher with a set of more flexible models. Bayesian models overcome the need to impose a priori sampling distributions on the efficiency term u. This approach treats the uncertainty in the choice of sampling model by mixing over a number of competing inefficiency distributions proposed in the literature with posterior model probabilities as weights. The choice of a particular distribution for the inefficiency term most favoured by the data can be made using Bayes factors or posterior odds ratio as a criterion for model selection. Bayesian models also allow parametric frontier modelling to deal with multiple outputs and undesirable outputs. Van den Broeck, Koop, Osiewalski & Steel (1994) first introduced Bayesian analysis in estimation of cross-sectional stochastic frontier models.

4.5 Nonparametric Techniques

Charnes, Cooper & Rhodes (1978) built on the pioneering work of Farrell (1957) by applying linear programming to estimate an empirical production technology frontier from which measures of efficiency can be obtained. The technique formally developed by Charnes et al. is known as Data Envelopment Analysis (DEA).

Data envelopment analysis is a mathematical programming model applied to observed data that allows construction of a production frontier as well as calculation of efficiency scores relative to the frontier. Based on the large and continually growing number of research papers published in this area, data envelopment analysis appears to be the popular choice for nonparametric efficiency analysis. The primary advantage of this technique is that there is no need to explicitly specify a mathematical form for the production function. However, because of its deterministic nature, it (usually) does not distinguish between technical inefficiency

and statistical noise.

It is important to note that the choice between a parametric approach or a nonparametric approach to efficiency analysis is not governed by which of these two approaches is superior; rather it should be determined by which approach is the most appropriate. For example, if there are only a small number of observations for analysis or if there is no satisfactory parameterisation of the frontier model, analysis should be directed towards a nonparametric approach. Charnes et al. (1994), Cooper et al. (2000) and Cooper et al. (2004) provide a comprehensive coverage on the theory and application of data envelopment analysis.

4.6 A Summary of Models and Estimation Techniques

Fried et al. (1993) and Kalirajan & Shand (1999) consider both nonparametric and parametric approaches to measuring productive and economic efficiency. Jacobs et al. (2006) considers both nonparametric and parametric approaches within a health care setting. Murillo-Zamorano (2004) provide a critical and detailed review of parametric and non-parametric frontier methods.

Figure 4.6 summarises the approaches to estimating production frontiers discussed in this chapter. For composed error $\varepsilon_i = v_i - u_i$ in the log-linear Cobb-Douglas (stochastic) production frontier model: $u_i = 0$ equates to a linear regression model; $v_i = 0$ and u_i only restricted by $u_i \geq 0$ equates to a deterministic production frontier model with parameters and technical efficiency being 'calculated'; $v_i = 0$ and $u_i \geq 0$ distributed asymmetrically equates to a deterministic production frontier model with parameters and technical efficiency being 'estimated'; and v_i distributed symmetrically and $u_i \geq 0$ distributed asymmetrically equates to a stochastic production frontier model.

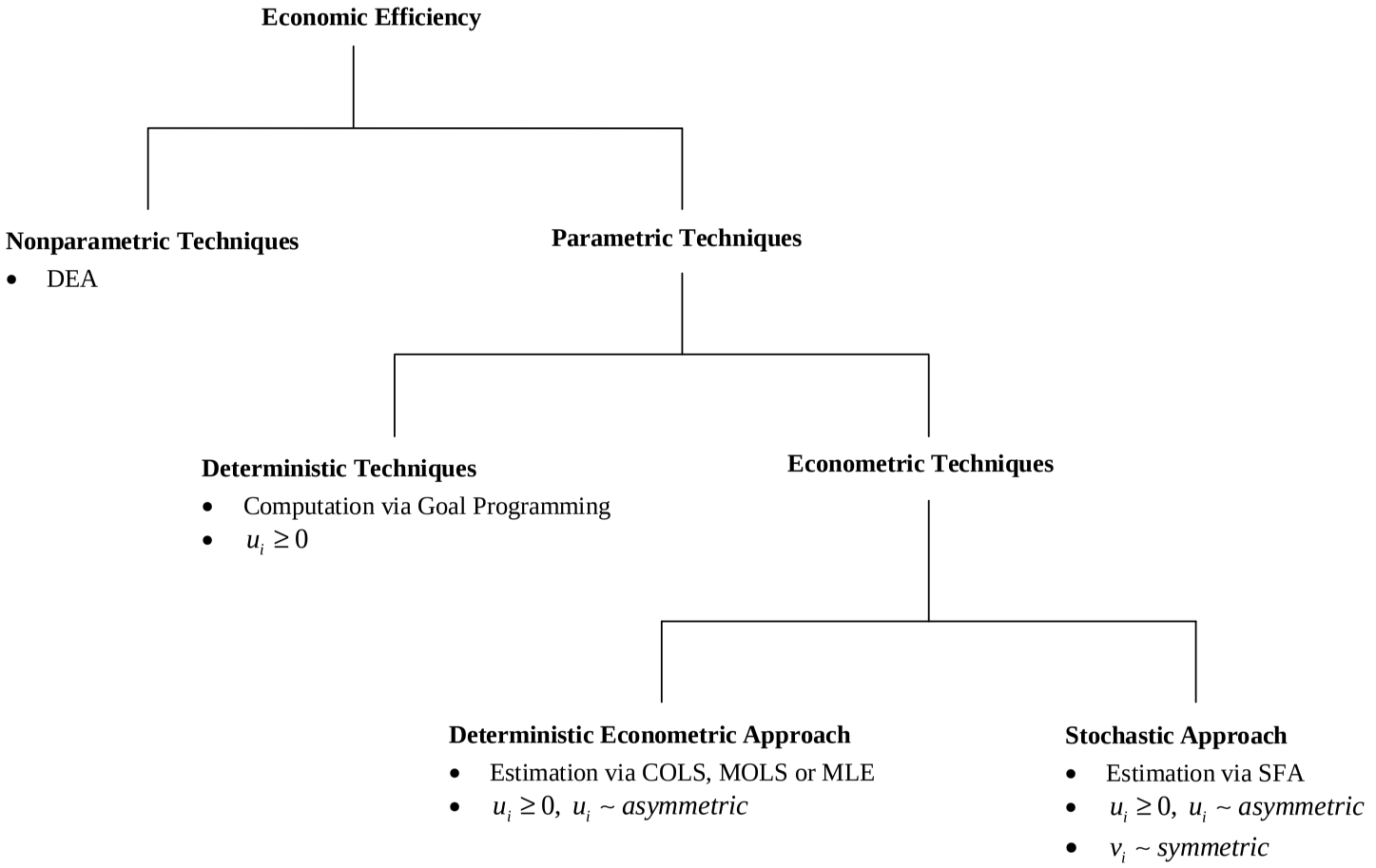

Figure 4.6: Estimation approaches for production frontier models.

Chapter 5

Optimum Design of Experiments

Consider the following model that was presented in Chapter 2

$$Y_i \;=\; f(\boldsymbol{x}_i, \boldsymbol{\beta}) + E_i, \qquad i = 1, \ldots, N. \tag{5.1}$$

A distributional assumption is imposed on the E_i with the assumed distribution having $k - p$ (possibly unknown) parameters $\boldsymbol{\tau}$. For example, if the E_i represent random error only, and not technical efficiency, with $E_i \sim N(0, \sigma^2)$ i.i.d. then $\boldsymbol{\tau}$ has just one element, which is σ^2. The true response $f(\boldsymbol{x}, \boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$, a vector of p unknown parameters that require estimation, and $\boldsymbol{x}$, a vector of m explanatory variables. Thus the full k-dimensional parameter vector $\boldsymbol{\theta}$ is partitioned into the p-dimensional parameter vector $\boldsymbol{\beta}$ from the model and the $(k-p)$-dimensional parameter vector $\boldsymbol{\tau}$ arising from the distributional assumption on the E_i, that is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$.

If the explanatory variables $\boldsymbol{x}$ can be controlled then experiments can be performed and the $\boldsymbol{x}$ are design variables with values belonging to a compact set $\mathscr{X}$ known as the design region or design space. The optimum design problem consists of choosing values for the design variables $\boldsymbol{x} \in \mathscr{X} \subseteq \mathbb{R}^m$ and determining the frequency or proportion of observations that should be taken at these values to optimise the estimation of the unknown parameters. Optimality is defined

using a criterion function Ψ which is to be maximimised. More will be discussed on criterion functions in later sections.

5.1 Linear Optimum Designs

Typically, linear optimum designs arise from linear models and nonlinear optimum designs arise from nonlinear models. The stochastic frontier model is an exception. The Cobb-Douglas form of a stochastic frontier is a linear model whose optimum design is nonlinear, that is, parameter dependent. For ease of comparison between linear and nonlinear optimum designs we consider the situation where an optimum linear design originates from a linear model.

Model (5.1) is linear when the i-th observation takes the form

$$Y_i \;=\; \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta} + E_i, \qquad i = 1, \ldots, N. \tag{5.2}$$

If the E_i have zero mean then the linear model can be expressed as

$$\mathbb{E}[Y] \;=\; F\boldsymbol{\beta},$$

where Y is the $N \times 1$ vector of responses and $F = [\boldsymbol{f}(\boldsymbol{x}_1), \ldots, \boldsymbol{f}(\boldsymbol{x}_N)]^T$ is an $N \times p$ matrix known as the model matrix. The i-th row of F is $\boldsymbol{f}^T(\boldsymbol{x}_i)$, a known function of the m explanatory variables. Additionally, assuming that the errors are independent with constant variance, the covariance matrix of the least squares estimate of $\boldsymbol{\beta}$ is

$$Cov(\hat{\boldsymbol{\beta}}) \;=\; \sigma^2 \left(F^T F\right)^{-1}.$$

The $p \times p$ matrix $F^T F$ is the information matrix of $\boldsymbol{\beta}$. The information in the experiment is greater for 'larger' values of $F^T F$. Hence the covariance of $\hat{\boldsymbol{\beta}}$ is 'smaller' for 'larger' $F^T F$. The advantage of least squares estimation is that it does not require any distributional assumption on the errors to estimate the

parameters. However, the least squares first order condition $\mathbb{E}[E_i] = 0$ may not hold, for example, for a stochastic frontier model. An alternative method of estimation is via maximum likelihood which assumes a distribution on the errors.

For the usual linear statistical model, the residuals are typically assumed to be independently and normally distributed with $E_i \sim N(0, \sigma^2)$, giving independently and normally distributed responses $Y_i \sim N(\boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}, \sigma^2)$. The probability density function of the i-th response is

$$f_{Y_i}(y_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left[y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}\right]^2\right\},$$

with log-likelihood function given by

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N}\left\{-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}[y_i - \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}]^2\right\}.$$

The expected Fisher information matrix of $\boldsymbol{\beta}$ derived from this log-likelihood function is

$$I_N(\boldsymbol{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^{N}\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) = \frac{1}{\sigma^2}F^TF, \tag{5.3}$$

giving the covariance matrix of the maximum likelihood estimate of $\boldsymbol{\beta}$ as

$$Cov(\tilde{\boldsymbol{\beta}}) = \{I_N(\boldsymbol{\beta})\}^{-1} = \sigma^2(F^TF)^{-1}.$$

Asymptotically, the variance covariance matrix of the maximum likelihood parameter estimates is the inverse of the Fisher information matrix. Hence designs that maximise the information in the experiment, minimise the variance of the estimates. The information matrix (5.3) is independent of the $\boldsymbol{\beta}$ parameters that require estimation but depends on σ^2, which may require estimation. Consequently, optimum designs for linear models are independent of the $\boldsymbol{\beta}$ parameters. Regardless of whether σ^2 is known or not, in comparing experimental designs for a specific experiment, the value of σ^2 is not relevant since the value is the same for all proposed designs.

If the parameter vector is extended to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ then the expected Fisher information matrix of $\boldsymbol{\theta}$ is the block diagonal matrix

$$I_N(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\left[\begin{array}{c|c} F^TF & \mathbf{0} \\ \hline \mathbf{0} & \dfrac{N}{2\sigma^2} \end{array}\right], \tag{5.4}$$

giving the covariance matrix of the maximum likelihood estimate of $\boldsymbol{\theta}$ as

$$Cov(\tilde{\boldsymbol{\theta}}) = \{I_N(\boldsymbol{\theta})\}^{-1} = \sigma^2\left[\begin{array}{c|c} (F^TF)^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \dfrac{2\sigma^2}{N} \end{array}\right].$$

The lower right element of information matrix (5.4) associated with the parameter σ^2 is independent of the design variables, thus we cannot design optimally for σ^2. In addition to this and the comments made above, the block diagonal structure of the information matrix implies that designs for optimal estimation of $\boldsymbol{\beta}$ are independent of σ^2, when comparing designs for a specific experiment. The optimum designs will be optimal for $\boldsymbol{\beta}$ with replicated design points required for estimation of σ^2.

The above example demonstrates some generalisations of linear optimum designs. The key point in comparing linear and nonlinear optimum designs is that linear optimum designs are independent of the $\boldsymbol{\beta}$ parameters that require estimation. The following section will demonstrate that this is not the case for nonlinear designs.

5.2 Nonlinear Optimum Designs

If the residuals in model (5.1) are independently and normally distributed with $E_i \sim N(0, \sigma^2)$ then the observed responses are independently and normally distributed with $Y_i \sim N(f(\boldsymbol{x}_i, \boldsymbol{\beta}), \sigma^2)$. The probability density function of the i-th response is

$$f_{Y_i}(y_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left[y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})\right]^2\right\},$$

with log-likelihood function given by

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; \sum_{i=1}^{N} \left\{ -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}[y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})]^2 \right\}.$$

The expected Fisher information matrix of $\boldsymbol{\beta}$ derived from this log-likelihood function is

$$I_N(\boldsymbol{\beta}) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{N} \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T, \tag{5.5}$$

giving the covariance matrix of the maximum likelihood estimate of $\boldsymbol{\beta}$ as

$$Cov(\tilde{\boldsymbol{\beta}}) \;=\; \{I_N(\boldsymbol{\beta})\}^{-1} \;=\; \sigma^2 \left[\sum_{i=1}^{N} \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \right]^{-1}.$$

Information matrix (5.5) may depend on the $\boldsymbol{\beta}$ parameters through the derivative of $f(\boldsymbol{x}_i, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Consequently, optimum designs for nonlinear models may depend on the unknown $\boldsymbol{\beta}$ parameters that require estimation.

If the parameter vector is extended to $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ then the expected Fisher information matrix of $\boldsymbol{\theta}$ is

$$I_N(\boldsymbol{\theta}) \;=\; \frac{1}{\sigma^2} \left[\begin{array}{c|c} \displaystyle\sum_{i=1}^{N} \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T & \mathbf{0} \\ \hline \mathbf{0} & \dfrac{N}{2\sigma^2} \end{array} \right]. \tag{5.6}$$

The block diagonal structure of the information matrix simplifies inversion giving the covariance matrix of the maximum likelihood estimate of $\boldsymbol{\theta}$ as

$$Cov(\tilde{\boldsymbol{\theta}}) \;=\; \{I_N(\boldsymbol{\theta})\}^{-1} \;=\; \sigma^2 \left[\begin{array}{c|c} \left[\displaystyle\sum_{i=1}^{N} \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial f(\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \right]^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \dfrac{2\sigma^2}{N} \end{array} \right].$$

The comments in the previous section regarding estimation of the parameter σ^2 apply with equal force here. Because the expected Fisher information matrix (5.6)

is block diagonal, as it is in equation (5.4) for the linear model, experiments can be designed for optimal estimation of $\boldsymbol{\beta}$ with replication required for estimation of σ^2. However, unlike linear optimum designs, nonlinear optimum designs may depend on the unknown $\boldsymbol{\beta}$ parameters. If prior values of the unknown parameters can be obtained from past experiments or studies then 'locally optimum' designs can be found. An alternative method is to impose prior distributions on the unknown parameters and obtain optimum Bayesian designs. Even if prior values of the unknown parameters are available, an optimum Bayesian design may be preferred over a locally optimum design as prior distributions on the parameters can reflect uncertainty in the prior values of the unknown parameters. Atkinson, Donev & Tobias (2007) discuss both locally optimum designs and optimum Bayesian designs for nonlinear models.

5.2.1 Nonlinear optimum designs for linear stochastic frontier models

The log-linear Cobb-Douglas form of the stochastic production frontier model (4.9), which was the focus of Chapter 4, can be expressed using linear model (5.2). Although the model is linear, optimum designs for estimating the unknown parameters are nonlinear due to the assumption of asymmetrically distributed errors. The same can be said for the more general form of the linear model (2.1) presented in Chapter 2.

For these linear models the overall error $E_i = c_u U_i + c_v V_i$, $\{c_u, c_v\} \in \mathbb{R}$, is composed of an asymmetrically distributed error term U_i and a symmetrically distributed random error term V_i giving a composed error E_i which is asymmetrically distributed. A result of the asymmetrical distribution of the composed error is that the E_i do not have zero mean. Consequently the information matrix of the parameters, and hence the covariance matrix of the estimates, are not

necessarily block diagonal. The expected Fisher information matrix given by

$$I_N(\boldsymbol{\theta}) = \mathbb{E}\left[\begin{array}{c|c} \left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\beta}}\right)^T & \left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\tau}}\right)^T \\ \hline \left\{\left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\tau}}\right)^T\right\}^T & \left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\tau}}\right)\left(\frac{\partial \ln \mathscr{L}}{\partial \boldsymbol{\tau}}\right)^T \end{array}\right],$$

or

$$I_N(\boldsymbol{\theta}) = -\mathbb{E}\left[\begin{array}{c|c} \frac{\partial^2 \ln \mathscr{L}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ln \mathscr{L}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\tau}^T} \\ \hline \frac{\partial^2 \ln \mathscr{L}}{\partial \boldsymbol{\tau}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 \ln \mathscr{L}}{\partial \boldsymbol{\tau}\partial \boldsymbol{\tau}^T} \end{array}\right],$$

for a log-linear stochastic production frontier model, is of the form

$$I_N(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\begin{array}{c|c} f_\beta(\boldsymbol{\tau})\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}_{\beta,\tau}^T(\boldsymbol{\tau}) \\ \hline \boldsymbol{f}_{\beta,\tau}(\boldsymbol{\tau})\boldsymbol{f}^T(\boldsymbol{x}_i) & F_\tau(\boldsymbol{\tau}) \end{array}\right]. \tag{5.7}$$

The function $\boldsymbol{f}^T(\boldsymbol{x}_i)$, which is a function of $\boldsymbol{x}_i$ only, is the i-th row of the model matrix and has length p. The function $f_\beta(\boldsymbol{\tau})$ is scalar-valued, and $\boldsymbol{f}_{\beta,\tau}(\boldsymbol{\tau})$ is a vector-valued function of length $k-p$. Both of the latter functions are functions of the $\boldsymbol{\tau}$ parameters and not the design variables $\boldsymbol{x}_i$. $F_\tau(\boldsymbol{\tau})$ is a symmetric matrix of dimension $k-p$ whose elements are functions of $\boldsymbol{\tau}$.

Unlike information matrices (5.4) and (5.6), information matrix (5.7) is not block diagonal, hence calculation of the covariance matrix of the maximum likelihood estimate of $\boldsymbol{\theta}$ is slightly more complicated. The covariance matrix is the inverse of the expected Fisher information matrix, that is $Cov(\tilde{\boldsymbol{\theta}}) = \{I_N(\boldsymbol{\theta})\}^{-1}$, where the information matrix is partitioned and Section E.1 in Appendix E gives the formulae for deriving the inverse of a partitioned matrix.

Information matrix (5.7) has a non-simple dependence on the $\boldsymbol{\tau}$ parameters only and does not depend on the $\boldsymbol{\beta}$ parameters. Consequently, optimum designs for log-linear stochastic frontier models depend on the $\boldsymbol{\tau}$ parameters and not the $\boldsymbol{\beta}$ parameters. Although model (5.2) is linear, an error term E_i with nonzero mean results in nonlinear parameter dependent optimum designs. The non-block diagonal structure of the information matrix implies that designs for optimal estimation of $\boldsymbol{\beta}$ may depend on the $\boldsymbol{\tau}$ parameters.

Optimum designs based on approximated information matrices

If the expected Fisher information matrix is approximated using, for example, the approximation methods described in Chapter 3, then it will have the form

$$\widehat{I}_N(\boldsymbol{\theta}) \;=\; \sum_{i=1}^{N} \left[\begin{array}{c|c} f_\beta(\mu_a)\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T_{\beta,\tau}(\mu_a) \\ \hline \boldsymbol{f}_{\beta,\tau}(\mu_a)\boldsymbol{f}^T(\boldsymbol{x}_i) & F_\tau(\mu_a) \end{array}\right],$$

where $\mu_a = \mathbb{E}[a_i]$ is a function of the $\boldsymbol{\tau}$ parameters only, and not the $\boldsymbol{\beta}$ parameters or the design variables $\boldsymbol{x}$. Hence, ultimately it has the same form and properties as the exact information matrix (5.7) that is not approximated.

5.3 Continuous and Exact Designs

Continuous designs are represented by the measure $\xi \in \Xi$ over $\mathscr{X}$ where Ξ is the class of all design measures on $\mathscr{X}$. Following the general notation of Atkinson et al. (2007), if the design has N trials at n ($n \leq N$) distinct points in $\mathscr{X}$, the process of determining an optimum design involves choosing a distribution over $\mathscr{X}$ written as

$$\xi \;=\; \left\{\begin{array}{cccc} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \\ w_1 & w_2 & \ldots & w_n \end{array}\right\}. \tag{5.8}$$

The first line gives the values of the design variables at the points of support of the design. The w_i are the associated design weights indicating the proportion of

observations that should be taken at each support point. Since ξ is a measure, $\int_{\mathscr{X}} \xi \, dx = \sum_{i=1}^{n} w_i = 1$ and $0 \leq w_i \leq 1$ for all i. The optimum continuous design is denoted by ξ^* with design points $\boldsymbol{x}_i^*$ and weights w_i^*. The information matrix for a continuous design is written

$$M(\xi) \;=\; \sum_{i=1}^{n} w_i \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) \;=\; F^T W F,$$

where

$$\begin{aligned} F^T &= \left[\boldsymbol{f}_\theta(\boldsymbol{x}_1), \ldots, \boldsymbol{f}_\theta(\boldsymbol{x}_n)\right], \\ W &= \operatorname{diag}\left(w_1, \ldots, w_n\right). \end{aligned}$$

The subscript θ in $\boldsymbol{f}_\theta(\boldsymbol{x}_i)$ is to emphasise that $M(\xi)$ is the information matrix of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ may be the extended parameter vector $(\boldsymbol{\beta}, \boldsymbol{\tau})$ and not just the $\boldsymbol{\beta}$ parameters.

If $I_i(\boldsymbol{\theta}) = \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$, the information matrix $M(\xi)$ is a weighted sum of per observation expected Fisher information matrices (c.f. Appendix D.2). The per observation expected Fisher information matrix $I_i(\boldsymbol{\theta})$ may not always be expressible in the form $\boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$; for example, when $I_i(\boldsymbol{\theta})$ is approximated using Method 2 or 3 in Sections 3.1.2 and 3.2.1 of Chapter 3 respectively. However the information matrix $M(\xi)$ can always be written as the (weighted sum of the) product of a column vector multiplied by its own transpose using an eigenvalue decomposition of the per observation expected Fisher information matrix. The eigendecomposition of a matrix with a structure like that given in Sections 3.1.2 or 3.2.1, where the information matrix is approximated, is detailed in Appendix D.5.

If the measure refers to an exact design, realisable in integer counts for a specific N, the measure is written as

$$\xi_N \;=\; \left\{ \begin{array}{cccc} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n \\ r_1/n & r_2/n & \ldots & r_n/N \end{array} \right\},$$

where r_i is the integer number of trials at $\boldsymbol{x}_i$ and $\sum_{i=1}^{n} r_i = N$. Exact designs can often be found by integer approximation to the optimum continuous design ξ^*, usually if the design weights w_i^* are rational. For exact N-trial designs the information matrix is

$$M(\xi_N) \;=\; \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) \;=\; \frac{1}{N}F^T F.$$

If $I_i(\boldsymbol{\theta}) = \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$ then $M(\xi_N) = \frac{1}{N}\sum_{i=1}^{N} I_i(\boldsymbol{\theta}) = \frac{1}{N} I_N(\boldsymbol{\theta})$ where $I_N(\boldsymbol{\theta})$ is the (full) expected Fisher information matrix (c.f. Appendix D.3).

Only continuous designs shall be considered in this dissertation, however, all designs used in practice are exact because an integer number of observations will be taken at each of the distinct points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$. For all the parameters that require estimation to be estimable, the number n of distinct design points must be at least as large as the number of parameters that require estimation. Thus for all the $\boldsymbol{\beta}$ parameters to be estimable $n \geq p$ and for the full parameter vector $\boldsymbol{\theta}$ to be estimable $n \geq k$. The general criterion function $\Psi = \Psi\{M(\xi)\}$, which is a function of the information matrix $M(\xi)$, is a general measure of precision. The design points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and their associated weights $w_1, \ldots, w_n$ are chosen to maximise Ψ, hence maximising precision and giving optimum estimates of the parameters.

5.4 Optimality Conditions

The relative merits of a design are typically determined using a scalar criterion function $\Psi\{M(\xi)\}$ which is a real-valued concave function defined on the class, $M(\Xi)$, of information matrices. The objective of optimum design is to

(i) maximise $\Psi\{M(\xi)\}$ over the set of information matrices $M(\xi) \in M(\Xi)$.

(ii) maximise $\Psi\{M(\xi)\}$ over the set of designs $\xi \in \Xi$.

(iii) maximise $\Psi\{M(\xi)\}$ over the set of design weights $\boldsymbol{w} \in \mathscr{W}$,

where $\mathscr{W} \equiv \{\boldsymbol{w} = (w_1, w_2, \ldots, w_n) : 0 \leq w_i \leq 1, \sum_{i=1}^{n} w_i = 1\}$. The above three objectives are equivalent. For the design measure ξ, the search for an optimum design usually involves finding the associated weights $\boldsymbol{w}$ for a set of fixed design points $\boldsymbol{x}$. As it is the design weights that ultimately define the optimum design, the objective as specified in (iii) shall be the focus for now. This is a nondegenerate constrained optimisation problem with the full constraint region being a closed bounded convex set. The criterion function can be written more explicitly as $\Psi(\boldsymbol{w})$ to emphasise its dependence on the weights. The conditions for an optimum design are defined in terms of directional derivatives of the criterion function Ψ.

5.4.1 Gâteaux directional derivative

Given a function $\Psi : \mathbb{R}^n \to \mathbb{R}$, the Gâteaux directional derivative of Ψ at $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ in the direction $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ is

$$G_\Psi(\boldsymbol{w}, \boldsymbol{v}) = \lim_{\alpha \to 0} \frac{\Psi(\boldsymbol{w} + \alpha\boldsymbol{v}) - \Psi(\boldsymbol{w})}{\alpha} = \frac{d}{d\alpha}\Psi(\boldsymbol{w} + \alpha\boldsymbol{v})\bigg|_{\alpha=0},$$

if the limit exists. If Ψ is differentiable

$$G_\Psi(\boldsymbol{w}, \boldsymbol{v}) = \left(\frac{d\Psi}{d\boldsymbol{w}}\right)^T \boldsymbol{v}.$$

The Gâteaux derivative is a more general form of the partial derivative, since if $\boldsymbol{v} = \boldsymbol{e}_i$

$$G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i) = \frac{\partial\Psi}{\partial w_i}\bigg|_{\boldsymbol{w}}.$$

The coordinate vector $\boldsymbol{e}_i \in \mathbb{R}^n$ has a 1 in the i-th position and zeros in the remaining $n - 1$ positions.

5.4.2 Fréchet directional derivative

The Fréchet directional derivative of Ψ at $\boldsymbol{w}$ in the direction $\boldsymbol{v}$ is

$$\begin{aligned} F_\Psi(\boldsymbol{w},\boldsymbol{v}) &= \lim_{\alpha\to 0}\frac{\Psi\left((1-\alpha)\boldsymbol{w}+\alpha\boldsymbol{v}\right)-\Psi(\boldsymbol{w})}{\alpha} \\ &= \lim_{\alpha\to 0}\frac{\Psi\left(\boldsymbol{w}+\alpha(\boldsymbol{v}-\boldsymbol{w})\right)-\Psi(\boldsymbol{w})}{\alpha} \\ &= G_\Psi(\boldsymbol{w},\boldsymbol{v}-\boldsymbol{w}). \end{aligned}$$

If the Gâteaux derivative is linear and if Ψ is differentiable then

$$\begin{aligned} F_\Psi(\boldsymbol{w},\boldsymbol{v}) &= G_\Psi(\boldsymbol{w},\boldsymbol{v})-G_\Psi(\boldsymbol{w},\boldsymbol{w}) \\ &= \left(\frac{d\Psi}{d\boldsymbol{w}}\right)^T\boldsymbol{v}-\left(\frac{d\Psi}{d\boldsymbol{w}}\right)^T\boldsymbol{w} \\ &= \sum_{j=1}^{n}\frac{\partial\Psi}{\partial w_j}v_j-\sum_{j=1}^{n}\frac{\partial\Psi}{\partial w_j}w_j. \end{aligned}$$

When $\boldsymbol{v}=\boldsymbol{e}_i$

$$\begin{aligned} F_\Psi(\boldsymbol{w},\boldsymbol{e}_i) &= G_\Psi(\boldsymbol{w},\boldsymbol{e}_i)-G_\Psi(\boldsymbol{w},\boldsymbol{w}) \\ &= \frac{d\Psi}{dw_i}-\left(\frac{d\Psi}{d\boldsymbol{w}}\right)^T\boldsymbol{w} \\ &= \frac{\partial\Psi}{\partial w_i}-\sum_{j=1}^{n}\frac{\partial\Psi}{\partial w_j}w_j. \end{aligned} \tag{5.9}$$

$F_\Psi(\boldsymbol{w},\boldsymbol{e}_i)$ is called the i-th vertex directional derivative of Ψ at $\boldsymbol{w}$ in the direction of the vertex $\boldsymbol{e}_i$. Assuming that $\Psi(\boldsymbol{w})$ is differentiable at $\boldsymbol{w}^*$, where $\boldsymbol{w}^*$ maximises $\Psi(\boldsymbol{w})$, the first-order conditions for a local maximum are

$$F_\Psi(\boldsymbol{w}^*,\boldsymbol{e}_i)\quad\begin{cases} = & 0 \quad \text{for } w_i^*>0 \\ \leq & 0 \quad \text{for } w_i^*=0. \end{cases} \tag{5.10}$$

This first-order stationarity condition is both necessary and sufficient for optimality.

When $\boldsymbol{w},\boldsymbol{v}\in\mathscr{W}$, they can be interpreted as distributions or sets of weights that define the design measure ξ. If ξ_i denotes the measure for the one point

design putting unit mass $w_i = 1$ at the point $\boldsymbol{x}_i$ then the weights for this measure are given by $\boldsymbol{e}_i$. Hence the directional derivative (5.9) of Ψ at $\boldsymbol{w}$ in the direction $\boldsymbol{e}_i$ can be equivalently represented as

$$\begin{aligned} F_\Psi\{M(\xi), M(\xi_i)\} &= G_\Psi\{M(\xi), M(\xi_i)\} - G_\Psi\{M(\xi), M(\xi)\} \\ &= \frac{\partial \Psi}{\partial w_i} - \sum_{j=1}^{n} \frac{\partial \Psi}{\partial w_j} w_j. \end{aligned}$$

By use of the chain rule, and for $M = M(\xi) = \sum_{i=1}^{n} w_i \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$,

$$\begin{aligned} F_\Psi\{M(\xi), M(\xi_i)\} &= \boldsymbol{f}_\theta^T(\boldsymbol{x}_i)\frac{\partial \Psi}{\partial M}\boldsymbol{f}_\theta(\boldsymbol{x}_i) - \sum_{j=1}^{n} \boldsymbol{f}_\theta^T(\boldsymbol{x}_j)\frac{\partial \Psi}{\partial M}\boldsymbol{f}_\theta(\boldsymbol{x}_j) w_j \\ &= \text{tr}\left\{\frac{\partial \Psi}{\partial M}\boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)\right\} - \text{tr}\left\{\frac{\partial \Psi}{\partial M}\sum_{j=1}^{n} w_j \boldsymbol{f}_\theta(\boldsymbol{x}_j)\boldsymbol{f}_\theta^T(\boldsymbol{x}_j)\right\} \\ &= \text{tr}\left\{\frac{\partial \Psi}{\partial M} M(\xi_i)\right\} - \text{tr}\left\{\frac{\partial \Psi}{\partial M} M(\xi)\right\}, \end{aligned}$$

where $M(\xi_i) = \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$ is the information matrix for the one point design putting unit mass at $\boldsymbol{x}_i$. That is, $F_\Psi\{M(\xi), M(\xi_i)\}$ is the directional derivative of Ψ at ξ in the direction of a one point design ξ_i. Atkinson et al. (2007) use this formulation of the directional derivative.

5.4.3 The General Equivalence Theorem

The following General Equivalence Theorem (Whittle 1973, White 1973, Kiefer 1974) provides alternative characterisations of an optimum design ξ^* such that $M(\xi^*)$ maximises $\Psi\{M(\xi)\}$.

Theorem 5.4.1 The following are equivalent

(i) $\Psi\{M(\xi)\}$ is maximised at $M(\xi^*)$.

(ii) $F_\Psi\{M(\xi^*), M(\xi)\} \leq 0$ for all $\xi \in \Xi$.

(iii) If Ψ is differentiable at $M(\xi^*)$ then $F_\Psi\{M(\xi^*), M(\xi)\}$ achieves its maximum at the one-point designs ξ_x, which put weight 1 at the support points ($\boldsymbol{x}$) of ξ^*.

The first-order condition (5.10) is an alternative, more concise, version of the General Equivalence Theorem, and moreover is now sufficient, as well as necessary, for global, not just local optimality, of $\boldsymbol{w}^*$. The General Equivalence Theorem holds for continuous designs but does not hold for exact designs in general. The theorem is useful for checking whether or not a proposed design is optimal and for motivating methods for the sequential construction of optimum designs. However it does not say anything about the value of n, the number of support points of the design.

The set $M(\Xi)$ is a convex and compact subset of the linear space, $\mathrm{Sym}(k)$, of symmetric matrices where the latter has dimension $\frac{1}{2}k(k+1)$. Hence a consequence of Carathéodory's Theorem is that most optimum designs are supported by at most $\frac{1}{2}k(k+1)$ points (Pukelsheim 1993). This is not true for Bayesian designs because the nonadditive nature of functions of information matrices precludes the use of Carathéodory's Theorem (Atkinson et al. 2007).

5.5 Optimality Criteria

The criterion for determining if a specific design is optimal or for comparing the optimality of different designs is based on the scalar-valued function $\Psi\{M(\xi)\}$, known as the criterion function, which is a function of the information matrix $M(\xi)$. Asymptotically, the variance covariance matrix of the parameter estimates is inversely proportional to the information matrix. Hence the covariance of the parameter estimates is smaller and the estimates are more precise in experiments with larger information. Maximisation of the criterion function corresponds to maximising information in the experiment. The criteria of op-

timality are often named using a letter of the alphabet so optimum design is sometimes referred to as 'alphabetic optimality'. The following are a selection of some of the most widely used optimality criteria and are the criteria of relevance for stochastic frontier models.

5.5.1 D-optimality

The most widely used design criterion is D-optimality where the log of the generalised variance is minimised so that the criterion function to be maximised is

$$\Psi\{M(\xi)\} \quad = \quad -\ln|M^{-1}(\xi)|.$$

The D-optimum design for a single parameter minimises the width of a confidence interval for the parameter of a linear model. For a two parameter linear model, D-optimum designs minimise the area of a confidence ellipse for the parameters. For a multidimensional parameter space, a D-optimum design minimises the volume of an ellipsoidal confidence region for the parameters of a linear model. Hence the criterion of D-optimality is used when interest is in estimating all k parameters as precisely as possible.

An advantage of D-optimality is that D-optimum designs are invariant to linear transformations, which is not generally the case for A-optimum designs. More on linear transformations is covered in the following chapter.

5.5.2 D_A-optimality

In D_A-optimality, interest is not in all k parameters, but only in s linear combinations of $\boldsymbol{\theta}$ which are the elements of $A^T\boldsymbol{\theta}$. The $k \times s$ matrix A has rank $s < k$. The covariance matrix for the s linear combination of the parameter estimates is given by

$$Cov(A^T\hat{\boldsymbol{\theta}}) \quad = \quad A^T M^{-1}(\xi)A.$$

D_A-optimality is an extension of D-optimality with criterion function

$$\Psi\{M(\xi)\} \;=\; -\ln|A^T M^{-1}(\xi)A|.$$

If $A^T = [I_s, \mathbf{0}]$, where I_s is the $s \times s$ identity matrix, then interest is in estimating a subset s of the parameters as precisely as possible. This is a special case of D_A-optimality known as *D_s-optimality*.

5.5.3 A-optimality

In A-optimality the average variance of the parameter estimates is minimised. This corresponds to maximising the negative of the average variance giving criterion function

$$\Psi\{M(\xi)\} \;=\; -\text{tr}\,\left\{M^{-1}(\xi)\right\}.$$

A-optimality is a special case of L-optimality.

5.5.4 L-optimality

For the $k \times q$ matrix L of coefficients, the criterion function to be maximised under linear (or L-) optimality is

$$\Psi\{M(\xi)\} \;=\; -\text{tr}\,\left\{M^{-1}(\xi)L\right\}.$$

If the rank of L is $s \leq q$ then it can be expressed as $L = AA^T$ where A is a $k \times s$ matrix of rank s, which gives criterion function

$$\Psi\{M(\xi)\} \;=\; -\text{tr}\,\left\{M^{-1}(\xi)AA^T\right\} \;=\; -\text{tr}\,\left\{A^T M^{-1}(\xi)A\right\}.$$

This is sometimes referred to as *A_A-optimality* and is an extension of A-optimality. When $s = 1$ the matrix A becomes the $k \times 1$ vector $\boldsymbol{c}$ and the optimality criterion is referred to as *c-optimality* with criterion function

$$\Psi\{M(\xi)\} \;=\; -\text{tr}\,\left\{\boldsymbol{c}^T M^{-1}(\xi)\boldsymbol{c}\right\} \;=\; -\boldsymbol{c}^T M^{-1}(\xi)\boldsymbol{c}.$$

Hence, a c-optimum design estimates the linear combination of the parameters $\boldsymbol{c}^T\boldsymbol{\beta}$ with minimum variance. If interest is in s linear combinations which are the elements of $C^T\boldsymbol{\theta}$, where C is a $k \times s$ matrix, then the criterion function is

$$\Psi\{M(\xi)\} \;=\; -\text{tr}\,\left\{C^T M^{-1}(\xi) C\right\},$$

and the criterion of optimality is known as *C-optimality.*

The above optimality criteria, their criterion functions and Gâteaux directional derivatives are summarised in Table 5.1. The functions appearing in this table can be used to calculate the Fréchet directional derivatives that appear in the General Equivalence Theorem 5.4.1. A general form for the Gâteaux derivative of Ψ at $M(\xi)$ in the direction of $M(\xi_i)$ is

$$G_\Psi\{M(\xi), M(\xi_i)\} \;=\; \boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} A \left\{A^T M^{-1} A\right\}^{(t-1)} A^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i).$$

For the D-criterion, $A = I_k$ and $t = 0$; $A = A$ and $t = 0$ for D_A- and D_s-optimality; $A = I_k$ and $t = 1$ for A-optimality; $A = \boldsymbol{c}$ and $t = 1$ for c-optimality; and for the C-criterion, $A = C$ and $t = 1$.

Criterion	Ψ	$G_\Psi\{M(\xi), M(\xi_i)\}$	$G_\Psi\{M(\xi), M(\xi)\}$
D	$-\ln \lvert M^{-1} \rvert$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i)$	k
D_A, D_s	$-\ln \lvert A^T M^{-1} A \rvert$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} A \left\{ A^T M^{-1} A \right\}^{-1} A^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i)$	$s = \text{rank } A$
A	$-\text{tr } \{M^{-1}\}$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i)$	$\text{tr}\, \{M^{-1}\}$
c	$-\boldsymbol{c}^T M^{-1} \boldsymbol{c}$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} \boldsymbol{c} \boldsymbol{c}^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i)$	$\boldsymbol{c}^T M^{-1} \boldsymbol{c}$
C	$-\text{tr } \left\{ C^T M^{-1} C \right\}$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} C C^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i)$	$\text{tr } \left\{ C^T M^{-1} C \right\}$

Table 5.1: Gâteaux derivatives appearing in the General Equivalence Theorem (5.4.1) for several optimality criteria.

5.5.5 Gâteaux derivatives for approximated information matrices

If the information matrix for a continuous design is approximated using Method 2 or 3 of Chapter 3 then it may be of the form

$$M(\xi) \;=\; \sum_{i=1}^{n} w_i \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C},$$

where $\mathscr{C}$ is a 'correction matrix' (c.f. Appendix D.5). In this case, the Gâteaux derivatives $G_\Psi\{M(\xi), M(\xi_i)\}$ in Table 5.1 are altered. Denote by $M_i = M(\xi_i)$, the information matrix for the one-point design ξ_i at $\boldsymbol{x}_i$

$$M(\xi_i) \;=\; \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C}.$$

If $M(\xi_i) = I_i(\boldsymbol{\theta})$ then it is the per observation expected Fisher information matrix of $\boldsymbol{\theta}$ at $\boldsymbol{x}_i$. For D_A- and D_s-optimality

$$\begin{aligned} G_\Psi\{M(\xi), M(\xi_i)\} &= \text{tr}\left\{M^{-1}A\left\{A^TM^{-1}A\right\}^{-1}A^TM^{-1}M_i\right\} \\ &= \text{tr}\left\{M^{-1}A\left\{A^TM^{-1}A\right\}^{-1}A^TM^{-1}\left[\boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C}\right]\right\} \\ &= \boldsymbol{f}_\theta^T(\boldsymbol{x}_i)M^{-1}A\left\{A^TM^{-1}A\right\}^{-1}A^TM^{-1}\boldsymbol{f}_\theta(\boldsymbol{x}_i) \\ &\qquad +\text{tr}\left\{M^{-1}A\left\{A^TM^{-1}A\right\}^{-1}A^TM^{-1}\mathscr{C}\right\}. \end{aligned}$$

The Gâteaux derivative for the D-criterion can be derived by letting $A = I_k$, where I_k is the $k \times k$ identity matrix. Similarly for C-optimality

$$\begin{aligned} G_\Psi\{M(\xi), M(\xi_i)\} &= \text{tr}\left\{M^{-1}CC^TM^{-1}M_i\right\} \\ &= \text{tr}\left\{M^{-1}CC^TM^{-1}\left[\boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C}\right]\right\} \\ &= \boldsymbol{f}_\theta^T(\boldsymbol{x}_i)M^{-1}CC^TM^{-1}\boldsymbol{f}_\theta(\boldsymbol{x}_i) + \text{tr}\left\{M^{-1}CC^TM^{-1}\mathscr{C}\right\}. \end{aligned}$$

If there is only one linear combination of interest then the $k \times s$ matrix C becomes the $k \times 1$ vector $\boldsymbol{c}$ and the appropriate Gâteaux derivative can be found by simply substituting $\boldsymbol{c}$ for C in the above equation. For A-optimality, the Gâteaux derivative $G_\Psi\{M(\xi), M(\xi_i)\}$ can be derived by substituting the matrix A for the matrix C above.

Criterion	Ψ	$G_\Psi\{M(\xi), M(\xi_i)\}$	$G_\Psi\{M(\xi), M(\xi)\}$
D	$-\ln \lvert M^{-1} \rvert$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i) + \text{tr}\ \{M^{-1}\mathscr{C}\}$	k
D_A, D_s	$-\ln \lvert A^T M^{-1} A \rvert$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} A \left\{A^T M^{-1} A\right\}^{-1} A^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i)$ $+\text{tr}\ \{M^{-1} A \left\{A^T M^{-1} A\right\}^{-1} A^T M^{-1}\mathscr{C}\}$	$s = \text{rank}\ A$
A	$-\text{tr}\ \{M^{-1}\}$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i) + \text{tr}\ \{M^{-1} M^{-1}\mathscr{C}\}$	$\text{tr}\ \{M^{-1}\}$
c	$-\boldsymbol{c}^T M^{-1} \boldsymbol{c}$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} \boldsymbol{c}\boldsymbol{c}^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i) + \text{tr}\ \{M^{-1}\boldsymbol{c}\boldsymbol{c}^T M^{-1}\mathscr{C}\}$	$\boldsymbol{c}^T M^{-1} \boldsymbol{c}$
C	$-\text{tr}\ \left\{C^T M^{-1} C\right\}$	$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) M^{-1} C C^T M^{-1} \boldsymbol{f}_\theta(\boldsymbol{x}_i) + \text{tr}\ \{M^{-1} C C^T M^{-1}\mathscr{C}\}$	$\text{tr}\ \left\{C^T M^{-1} C\right\}$

Table 5.2: Gâteaux derivatives, for approximated information matrix $M(\xi)$, appearing in the General Equivalence Theorem (5.4.1) for several optimality criteria.

For optimality criteria in general, the 'corrected' Gâteaux derivative $G_\Psi\{M(\xi), M(\xi_i)\}$ is composed of two terms; the first term has the form of the Gâteaux derivative $G_\Psi\{M(\xi), M(\xi_i)\}$ for the exact information matrix, like those that appear in Table 5.1, although it is *not* the same because $\boldsymbol{f}_\theta(\boldsymbol{x}_i)$ will be different; the second term is a 'correction' to the derivative. Table 5.2 parallels Table 5.1 and summarises the Gâteaux directional derivatives for the aforementioned optimality criteria for approximated $M(\xi)$. Note that the derivative $G_\Psi\{M(\xi), M(\xi)\}$ remains unchanged.

5.6 Optimum Design Measures with Singular Information Matrices

If $M(\xi^*)$ is singular, for example when the number of support points n is less than the number of parameters k to be estimated, only certain linear combinations or subsets of the parameters may be estimable. Silvey (1980) considers optimum design measures with singular information matrices and provides a sufficient condition for a design measure with singular information matrix to be optimal, for both linear and nonlinear models.

Let $M(\Xi)$ be the set of information matrices generated as ξ ranges over the set of all distributions Ξ on $\mathscr{X}$. $M(\Xi)$ is a convex set of nonnegative definite matrices. A typical element of $M(\Xi)$ is denoted by $M(\xi)$. By suppressing the argument ξ, $M(\xi)$ is denoted simply by M.

Suppose that interest is in certain linear combinations of the unknown parameters, say the vector $A^T\boldsymbol{\theta}$, where A is a $k \times s$ matrix of rank $s < k$. A design measure ξ with information matrix M allows estimation of $A^T\boldsymbol{\theta}$ if $M\boldsymbol{z} = \mathbf{0}$ implies $A^T\boldsymbol{z} = \mathbf{0}$, that is, the null space of M is contained in that of A, or equivalently if $A = MY$, for some matrix Y. Let $M_A(\Xi)$ be the subset of $M(\Xi)$ consisting of those M with this property. It is assumed that $\mathscr{X}$ is such that $M_A(\Xi)$ is

nonempty. Some $M \in M_A(\Xi)$ are singular which can cause considerable problems in optimum design theory.

The covariance matrix of the least squares estimator of $A^T\boldsymbol{\theta}$ arising from a design measure with information matrix $M \in M_A(\Xi)$ is proportional to $A^T M^- A$, where M^- is *any* generalised inverse of M, that is any matrix such that $MM^-M = M$. Note that $A^T M^- A$ does not depend on which generalised inverse is chosen; also that $A^T M^- A$ is a positive definite $s \times s$ matrix. Interest in $A^T\boldsymbol{\theta}$ implies that the design objective will be to make some function of $A^T M^- A$ small in some sense.

Typically an optimum design will aim to maximise a criterion function $\Psi\{M(\xi)\}$ defined by

$$\Psi\{M(\xi)\} = \begin{cases} \varphi(A^T M^- A), & M \in M_A(\Xi), \\ -\infty, & \text{otherwise.} \end{cases} \tag{5.11}$$

Here φ is a real-valued function that is finite on the positive definite $s \times s$ matrices. It is assumed that Ψ is convex on $M_A(\Xi)$; also that Ψ is differentiable at nonsingular M, when $M^- = M^{-1}$, and nondifferentiable at singular M, but the directional derivatives can still be defined. With regards to Table 5.1, if M is singular, the generalised inverse M^- should replace M^{-1} for D_A-, D_s-, c- and C-optimality.

Silvey (1978) establishes a sufficient condition for a convex criterion function of the form (5.11) to be maximised by a design measure with singular information matrix. Extending the work of Silvey (1978), Ford & Silvey (1980) investigate the properties of a design constructed sequentially for a simple nonlinear problem where the optimum design measure has singular information matrix. For regression with uncorrelated observations, Fedorov (1978) gives a test for optimality using the generalised inverse of M when the matrix for design, M, is degenerate. If the experimental design region is augmented, as in extended experiments, Pázman (1978) proposes that generalised inverses are not needed for computing

optimum designs when the singularity of the information matrix in unavoidable; however they are needed for the analysis. Because interest here is not in extended experiments the problem of which generalised inverse of M to use in calculating $A^T M^- A$ requires some consideration.

5.6.1 Generalised inverses

The nomenclature for the various types of generalised inverses are not standard. Rohde (1965) gives the following definitions for several generalised inverses.

Definition 5.6.1 The *generalised inverse* of a matrix X is a matrix $X^{(g)}$ such that

$$XX^{(g)}X = X.$$

Definition 5.6.2 $X^{(r)}$ denotes a *reflexive generalised inverse* or a *semi-inverse* if it obeys the relations

$$\begin{aligned} XX^{(r)}X &= X, \\ X^{(r)}XX^{(r)} &= X^{(r)}. \end{aligned}$$

Definition 5.6.3 $X^{(N)}$ denotes a *normalised generalised inverse* or a *weak generalised inverse* if it obeys the relations

$$\begin{aligned} XX^{(N)}X &= X, \\ X^{(N)}XX^{(N)} &= X^{(N)}, \\ XX^{(N)} &= [XX^{(N)}]^H, \end{aligned}$$

i.e. $XX^{(N)}$ is Hermitian.

Definition 5.6.4 $X^\dagger$ denotes a *pseudoinverse* or a *Moore-Penrose generalised inverse* if it obeys the relations

$$\begin{aligned} XX^\dagger X &= X, \\ X^\dagger XX^\dagger &= X^\dagger, \\ XX^\dagger &= (XX^\dagger)^H, \\ (X^\dagger X) &= (X^\dagger X)^H, \end{aligned}$$

i.e. $XX^\dagger$ and $X^\dagger X$ are Hermitian. The pseudoinverse, which is uniquely determined by X, was independently described by Moore (1920) and later by Penrose (1955) under the names general reciprocal and generalised inverse, respectively.

Here A^H is the conjugate transpose of a matrix A. A Hermitian matrix is a square matrix with complex entries equal to its own conjugate transpose. For matrices whose elements are real numbers, rather than complex numbers, $A^H = A^T$. If X is square and non-singular, the above defined generalised inverses reduce to X^{-1}.

A more general nomenclature for the various types of generalised inverses is defined by Ben-Israel & Greville (1974) using the following equations:

(1) $XX^-X = X$,

(2) $X^-XX^- = X^-$,

(3) $XX^- = (XX^-)^H$, i.e. XX^- is Hermitian

(4) $X^-X = (X^-X)^H$, i.e. X^-X is Hermitian.

Definition 5.6.5 In general, if a matrix X^- satisfies equations (i), (j), and (k), then X^- is called an (i,j,k)-inverse of X, i.e. the generalised inverse is a (1)-inverse, $X^- = X^{(g)}$; the reflexive generalised inverse is a (1,2)-inverse, $X^- = X^{(r)}$; the normalised generalised inverse is a (1,2,3)-inverse, $X^- = X^{(N)}$; the pseudoinverse is the (1,2,3,4)-inverse, $X^- = X^\dagger$.

It is clear from these definitions that the various types of generalised inverses are, in general, not equivalent. The (1,2,3,4)-inverse or the pseudoinverse $M^- = M^\dagger$ of a matrix M is the only generalised inverse that is uniquely determined by M. For this reason, the pseudoinverse will be used in calculating $A^T M^- A$ for those optimum design measures with singular information matrices.

From Rohde (1965), if M is a nonnegative definite Hermitian matrix then we can write

$$M = [X_1|X_2]^H[X_1|X_2] = \left[\begin{array}{c|c} A & C \\ \hline C^H & B \end{array}\right],$$

where $A = X_1^H X_1$, $C = X_1^H X_2$, $B = X_2^H X_2$. A generalised inverse of matrix M is a matrix

$$M^{(g)} = \left[\begin{array}{c|c} A^{(g)} + A^{(g)}CQ^{(g)}C^H A^{(g)} & -A^{(g)}CQ^{(g)} \\ \hline -Q^{(g)}C^H A^{(g)} & Q^{(g)} \end{array}\right], \tag{5.12}$$

where $Q = B - C^H A^{(g)} C$.

Lemma 5.6.1 If the nonnegative $k \times k$ Hermitian matrix M is partitioned as

$$M = \left[\begin{array}{c|c} A & C \\ \hline C^H & B \end{array}\right],$$

where A is $(k-q) \times (k-q)$ of rank p, B is $q \times q$ of rank q, and M is of rank $p+q$, then

$$Q = B - C^H A^{(g)} C$$

is nonsingular.

Theorem 5.6.1 If a nonnegative Hermitian matrix M is partitioned in the form

$$M = \left[\begin{array}{c|c} A & C \\ \hline C^H & B \end{array}\right],$$

then

(i) a generalised inverse of M is given by (5.12),

(ii) a reflexive generalised inverse of M is given by (5.12) with $A^{(g)}$ and $Q^{(g)}$ replaced by $A^{(r)}$ and $Q^{(r)}$.

Further if rank M = rank A + rank B, where B is nonsingular, then

(iii) a normalised generalised inverse of M is given by (5.12) with $A^{(g)}$ and $Q^{(g)}$ replaced by $A^{(N)}$ and $Q^{(N)}$,

(iv) a pseudoinverse of M is given by (5.12) with $A^{(g)}$ and $Q^{(g)}$ replaced by $A^{\dagger}$ and $Q^{\dagger}$.

Rohde (1966) also gives the following useful theorem.

Theorem 5.6.2 A necessary and sufficient condition that

$$\text{rank } X \;=\; \text{rank } X^{(g)}$$

is that $X^{(g)}$ be a reflexive generalised inverse of X.

Pseudoinverse of the information matrix $M(\xi)$

It was noted earlier that if X is square and nonsingular, the generalised inverses reduce to X^{-1}. That is, for the pseudo- or Moore-Penrose generalised inverse, if the inverse of X^TX exists then

$$X^{-} \;=\; (X^TX)^{-1}X^T.$$

In general, for X and Y square, $(XY)^{-} \neq Y^{-}X^{-}$ unless X and Y are of full rank, i.e. rank X = rank Y = number of rows or columns in X or Y.

In Section 5.3, it was shown that the information matrix for any continuous design can be represented as

$$M(\xi) \;=\; F^TWF$$

where F is an $n \times k$ matrix and W is a diagonal matrix of dimension n. If the number of distinct design points n is less than the number of parameters k to be estimated then $M = M(\xi)$ will be singular. M may also be singular if the information matrix is approximated. For example, under approximation Method 1 of Chapter 3, a $k \times k$ information matrix will have at most rank $\leq (p+1)$, $p < k$. The ranks of the constituent matrices are

$$\text{rank } F \;=\; \text{rank } F^T \;=\; \text{rank } W \;=\; n,$$

hence

$$\begin{aligned} M^- &= F^- W^- (F^-)^T \\ &= F^- W^{-1} (F^-)^T, \end{aligned}$$

where $W^- = W^{-1}$ since W is a nonsingular diagonal matrix. Now, the pseudoinverse of F is

$$F^- \;=\; (F^T F)^{-1} F^T,$$

which yields the pseudoinverse

$$M^- \;=\; (F^T F)^{-1} F^T W^{-1} \left\{ (F^T F)^{-1} F^T \right\}^T.$$

If the information matrix is approximated, using Methods 2 or 3 from Chapter 3 for example, then from Appendix D.5, it can be expressed as

$$M(\xi) \;=\; F^T W F + Q_{22} \Lambda_{22} Q_{22}^T \;=\; R^T S R,$$

where R is an $(n + k - p) \times k$ matrix and S is a diagonal matrix of dimension $(n + k - p)$. For singular M, similar calculations to those carried out above give

$$M^- \;=\; (R^T R)^{-1} R^T S^{-1} \left\{ (R^T R)^{-1} R^T \right\}^T.$$

5.7 A Multiplicative Algorithm for Constructing Optimising Distributions

It was noted earlier that the General Equivalence Theorem can be used to motivate the construction of optimum designs and to verify if the proposed design is optimal. We shall consider the General Equivalence Theorem as given by equation (5.10)

$$F_\Psi(\boldsymbol{w}^*, \boldsymbol{e}_i) \quad \begin{cases} = & 0 \quad \text{for } w_i^* > 0 \\ \leq & 0 \quad \text{for } w_i^* = 0, \end{cases}$$

where the Fréchet derivative $F_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)$ is the i-th vertex directional derivative of Ψ at $\boldsymbol{w}$ in the direction $\boldsymbol{e}_i$. From equation (5.9) the Fréchet directional derivative is

$$\begin{aligned} F_\Psi(\boldsymbol{w}, \boldsymbol{e}_i) &= \frac{\partial \Psi}{\partial w_i} - \sum_{j=1}^{n} w_j \frac{\partial \Psi}{\partial w_j} \\ &= G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i) - \sum_{j=1}^{n} w_j G_\Psi(\boldsymbol{w}, \boldsymbol{e}_j), \end{aligned}$$

where the partial derivative $\dfrac{\partial \Psi}{\partial w_i} = G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)$ is the Gâteaux directional derivative of Ψ at $\boldsymbol{w}$ in the direction $\boldsymbol{e}_i$.

A multiplicative algorithm, first proposed by Torsney (1977), for obtaining the optimum design weights $w_i^* \in \mathscr{W}$ is

$$w_i^{(k+1)} = \frac{w_i^{(k)} f\left\{G_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_i)\right\}}{\sum_{i=1}^{n} w_j^{(k)} f\left\{G_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_j)\right\}}, \tag{5.13}$$

where $G_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_i) = \left.\dfrac{\partial \Psi}{\partial w_i}\right|_{w_i = w_i^{(k)}}$ and $f\left\{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\right\}$ is a positive and strictly increasing function of the derivative $G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)$. The notation $w_i^{(k)}$ denotes the value of the i-th weight at iteration k. The numerator $\sum w_j^{(k)} f\left\{G_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_j)\right\}$ is a scaling factor that ensures that $\sum w_i^{(k+1)} = 1$. A suitable choice of weights to

initiate the algorithm would be to choose $w_i^{(0)} = 1/n$ and a suitable stopping criteria for termination of the algorithm is $\max_{1 \leq i \leq n} \{F_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\} \leq 10^{-n}$. Appendix F.1 provides further details on implementation of the algorithm including some pseudocode.

5.7.1 Properties of the iteration

Torsney's multiplicative algorithm obeys the following four properties, which can be useful in monitoring the convergence of the iterations.

Property 1 $w^{(k)}$ is always feasible.

Property 2 $F_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{w}^{(k+1)}) \geq 0$ with equality when the $\dfrac{\partial \Psi}{\partial w_i}$ corresponding to nonzero w_i are all equal (in which case $w^{(k+1)} = w^{(k)}$). Recall that $G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i) = \dfrac{\partial \Psi}{\partial w_i}$.

Property 3 $supp(\boldsymbol{w}^{(k+1)}) \equiv supp(\boldsymbol{w}^{(k)})$ but weights can converge to zero. That is, the points of support of the design remain the same for all iterates but their associated weights can converge to zero.

Property 4 An iterate $w^{(k)}$ is a fixed point of the iteration if the derivatives $\dfrac{\partial \Psi}{\partial w_i^{(k)}}$ corresponding to nonzero $w_i^{(k)}$ are all equal. This is a necessary but not a sufficient condition for $w_i^{(k)}$ to maximise the criterion function.

Torsney (1977) implemented the following function of the derivative

$$f\{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\} \;=\; \{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\}^\delta, \qquad \delta \in \mathbb{R}^+,$$

and reports that the best choices of δ for the determinant and trace criterion respectively are $\delta = 1$ and $\delta = 0.5$. The choice of $\delta = 1$ for the determinant criterion is based on the results of Baum & Eagon (1967), and that $\delta = 0.5$ for

the trace criterion was guided by the results of Fellman (1974). Because $\delta > 0$, the derivatives $G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)$ are required to be positive.

Other choices of $f(\cdot)$ and δ have been considered. Silvey, Titterington & Torsney (1978) explore choices of δ for Torsney's (1977) initial function $f\{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\} = \{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\}^\delta$ and propose choosing δ on an ad hoc basis for both a nonadaptive and an adaptive algorithm. Torsney (1988) considers $f\{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\} = \exp\{\delta G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\}$ with applications in design, estimation and image processing. Torsney & Alahmadi (1992) contribute further algorithmic developments for the multiplicative algorithm above. Mandal & Torsney (2000) investigate the use of $f\{G_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\}$ and $f\{F_\Psi(\boldsymbol{w}, \boldsymbol{e}_i)\}$, for various choices of $f(\cdot)$. Torsney & Mandal (2001) extend Alahmadi's (1993) earlier work in which the constrained optimisation problem is transformed to one of simultaneous maximisation of two objective functions with respect to design weights. In an attempt to improve convergence, Torsney & Mandal (2004) suggest objective choices of $f(\cdot)$ which allows the criterion function to have negative derivatives. Mandal & Torsney (2006) consider developments of the above algorithm based on a clustering approach motivated by the practical application of the algorithm often giving the optimum design as a distribution defined on a disjoint cluster of points. Fedorov (1972) and Wynn (1972) also consider vertex direction algorithms which perturb one weight and change the others proportionately.

5.8 Further Reading

For further reading on optimum experimental design Atkinson & Donev (1992) provide good coverage on the topic. Their 2007 Atkinson et al. text is essentially their Atkinson & Donev (1992) text with some additional material, the inclusion of SAS[1] examples and coauthored by an additional author, Randall

[1]The SAS System (originally *Statistical Analysis System*) is an integrated system of software products provided by SAS Institute that enables the programmer to carry out statistical tasks.

Tobias. Melas's (2006) text focusses on a functional approach to optimum experimental design. Fedorov (1972), Silvey (1980) and Pukelsheim (1993) are also classic texts in optimum design.

Some articles whose nature is more of a review of optimum design include: Wynn's (1984) article summarising Jack Kiefer's contributions to experimental design; Atkinson & Fedorov (1989) appears in the supplement of the *Encyclopedia of Statistical Sciences* and provides a brief overview of optimum design; Atkinson (1996) discusses the usefulness of optimum experimental designs; and Atkinson & Bailey (2001) give a summary of design articles appearing in *Biometrika* over a 100 year time span. Fedorov & Läuter (1987), Dodge, Fedorov & Wynn (1988), Atkinson, Bogacka & Zhigljavsky (2001) and Di Bucchianico, Läuter & Wynn (2004) are conference proceedings containing useful articles on optimum design.

Areas of optimum design not covered in this dissertation, but that are gaining considerable popularity due to current environmental issues, are optimum spatial design and optimum Bayesian design. Müller (2001) is an excellent introductory text on optimum designs for spatial data and Chaloner & Verdinelli (1995) review Bayesian experimental designs.

It is somewhat of an annoyance that the mathematical notation used throughout the literature is inconsistent, particularly prior to Atkinson & Donev's 1992 edition of their text. However, many author's have since adopted notation generally consistent with that in Atkinson & Donev's (1992) text.

Chapter 6

Optimum Designs for Stochastic Production Frontier Models

The log-linear Cobb-Douglas form of the single-output stochastic production frontier model given in equation (4.9) of Chapter 4 for the i-th observational unit is

$$\ln y_i \;=\; \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_{ij} + v_i - u_i,$$

with expected log output given by

$$\mathbb{E}[\ln Y_i] \;=\; \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_{ij} - \mathbb{E}[U_i],$$

since random error $V \sim N(0, \sigma_v^2)$ i.i.d. A logarithmic transform is applied to the response y and the predictors x_j in a log-linear stochastic frontier model. To simplify notation, the logarithms will not be written explicitly in further models but it should be noted that a logarithmic transformation of the inputs and outputs should be carried out prior to estimation of the parameters.

We shall consider only models where technical efficiency, represented by U, is half normally or exponentially distributed. This is motivated by Ritter & Simar's (1997) argument that the one-parameter half normal and exponential

specifications should be preferred over the two-parameter truncated normal and gamma distributions because the resultant efficiency rankings are not sensitive to distributional assumptions. The information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \sigma_G^2)$ for a model with a normal-half normal error specification is given in Section 4.3.1 where $\lambda = \sigma_u/\sigma_v$, $\sigma_G^2 = \sigma_u^2 + \sigma_v^2$ and σ_u^2 is the parameter from the half normal distribution. The information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, 1/\sigma_u, \sigma_v^2)$ for a model with a normal-exponential error specification is given in Section 4.3.2 where $1/\sigma_u$ is the rate parameter from the exponential distribution. The parameter vector $\boldsymbol{\theta}$ can be written more generally as $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ where $\boldsymbol{\tau} = (\lambda, \sigma_G^2)$ for half normally distributed efficiency and $\boldsymbol{\tau} = (1/\sigma_u, \sigma_v^2)$ for exponentially distributed efficiency.

The elements of the information matrix for both error specifications involve complicated expectations, hence the recommended approximation method, Method 1, from Chapter 3 shall be used to calculate approximated information matrices in any numerical examples. Using this method, the first-order derivatives of the log-likelihood function are approximated by a first-order Taylor polynomial in the covariance definition of the expected Fisher information matrix. For a log-linear stochastic production frontier model, the approximated information matrix is of the form

$$M(\xi) = \left[\begin{array}{c|c} f_\beta(\mu_a)^2 \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \sum w_i \boldsymbol{f}(\boldsymbol{x}_i) f_\beta(\mu_a)\boldsymbol{f}_\tau^T(\mu_a) \\ \hline \boldsymbol{f}_\tau(\mu_a) f_\beta(\mu_a) \sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i) & \boldsymbol{f}_\tau(\mu_a)\boldsymbol{f}_\tau^T(\mu_a) \end{array}\right], \quad (6.1)$$

where μ_a is a function of the $\boldsymbol{\tau}$ parameters only, and not the $\boldsymbol{\beta}$ parameters or the design variables $\boldsymbol{x}$. Properties of these approximated information matrices are discussed in Section 3.1.1.

If Method 2 or Method 3 from Chapter 3 are used to approximate the information matrix, the information matrix is not guaranteed to be nonnegative definite, although the information matrix can be perturbed to ensure nonsigular-

ity (Atkinson et al. 2007). For a log-linear stochastic production frontier model, the approximated information matrix is of the form

$$M(\xi) \;=\; \left[\begin{array}{c|c} f_\beta(\mu_a)\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T_{\beta,\tau}(\mu_a) \\ \hline \boldsymbol{f}_{\beta,\tau}(\mu_a)\sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i) & F_\tau(\mu_a) \end{array}\right]. \qquad (6.2)$$

Under approximation Methods 2 and 3, the Gâteaux derivative $G_\Psi\{M(\xi), M(\xi_i)\}$, used in multiplicative algorithm (5.13) for finding optimising design weights, requires a 'correction'. The correction to the derivative is discussed in Section 5.5.5.

Some general properties of optimum designs for stochastic frontier models are discussed briefly in Section 5.2.1. The designs may be parameter dependent due to a non-simple dependence on the $\boldsymbol{\tau}$ parameters. Additionally, the information matrix is singular. Consequently, only subsets or linear combinations of the parameters can be estimated. Results on linear transformations of the design space and parameter space are reviewed in Section 6.1. Admissible transformations for the stochastic frontier model, which addresses the singularity of the information matrix, are then presented in Section 6.2. Section 6.3 equates transformations on a p-dimensional parameter space with transformations on a $(k > p)$-dimensional parameter space for frontier models. Finally, theoretical and numerical results for the determinant and trace criterion functions are presented in Sections 6.4 and 6.5 respectively.

6.1 Linear Transformations

It was noted above that, for log-linear stochastic production frontier models, a logarithmic transformation of the response y and explanatory variables x_j should be applied prior to estimation of the parameters. Some design criteria are not invariant to linear transformations. The following considers the effects of linear

transformations on determinant and trace criterion functions for linear models.

6.1.1 Linear transformation of the design space

Let the information matrix be denoted by

$$M_f = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i).$$

A linear transformation of the design space can be expressed as

$$\boldsymbol{g}(\boldsymbol{x}_i) = A\boldsymbol{f}(\boldsymbol{x}_i), \tag{6.3}$$

with inverse transformation

$$\boldsymbol{f}(\boldsymbol{x}_i) = A^{-1}\boldsymbol{g}(\boldsymbol{x}_i) = B\boldsymbol{g}(\boldsymbol{x}_i),$$

where A is a $k \times k$ matrix and $|A| \neq 0$. The information matrix can then be re-expressed as

$$\begin{aligned} M_f &= \sum w_i B\boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}^T(\boldsymbol{x}_i)B^T \\ &= B\left(\sum w_i \boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}^T(\boldsymbol{x}_i)\right)B^T \\ &= BM_gB^T, \end{aligned}$$

where

$$M_g = \sum w_i \boldsymbol{g}(\boldsymbol{x}_i)\boldsymbol{g}^T(\boldsymbol{x}_i).$$

Determinant criterion functions

Theorem 6.1.1 *D*-optimum designs are invariant to linear transformations of the design space.

Proof The criterion function for *D*-optimality is given by

$$\begin{aligned} -\ln|M_f^{-1}| &= -\ln|(B^T)^{-1}M_g^{-1}B^{-1}| \\ &= -\ln|(B^T)^{-1}| - \ln|M_g^{-1}| - \ln|B^{-1}| \\ &= -\ln|B^{-1}|^2 - \ln|M_g^{-1}|. \end{aligned}$$

The derivative with respect to the design weights is

$$-\frac{d}{d\boldsymbol{w}} \ln |M_f^{-1}| \;=\; -\frac{d}{d\boldsymbol{w}} \ln |M_g^{-1}|,$$

since $B^{-1} = A$ is a matrix that is independent of the design weights. That is, the optimum design which maximises $\Psi\{M_f(\xi)\}$ also maximises $\Psi\{M_g(\xi)\}$. Hence D-optimum designs are invariant to linear transformations of the design space. □

It is a straightforward extension to show that D_A-optimum designs are also invariant to linear transformations of the design space using

$$A^T M_f^{-1} A \;=\; A(B^T)^{-1} M_g^{-1} B^{-1} A \;=\; \tilde{A}^T M_g^{-1} \tilde{A},$$

where $\tilde{A} = B^{-1}A$. Hence D_A-optimality transforms to $D_{\tilde{A}}$-optimality.

Corollary 6.1.1 It follows from Theorem 6.1.1 that, D-optimum designs on the scaled design space $[ab, b]$ can be calculated from the optimum designs on the interval $[a, 1]$ by multiplying the support points by b (Chang 1999). Hence D-optimum designs are scale invariant but not necessarily translation invariant.

Trace criterion functions

Theorem 6.1.2 *A*-optimum designs are not necessarily invariant to linear transformations of the design space.

Proof The criterion function for A-optimality is given by

$$-\text{tr}\,\{M_f^{-1}\} \;=\; -\text{tr}\,\{(B^T)^{-1} M_g^{-1} B^{-1}\}.$$

In general, this is not proportional to $-\text{tr}\,\{M_g^{-1}\}$, therefore, generally

$$-\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M_f^{-1}\} \;\neq\; -\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M_g^{-1}\}.$$

For the A-criterion, the design that maximises $\Psi\{M_f(\xi)\}$ is not necessarily the design that maximises $\Psi\{M_g(\xi)\}$ on the transformed design space. □

Similar arguments apply in showing that the more general C- or L-optimum designs are not necessarily invariant to linear transformations of the design space. However, both A- and the more general L-criteria are 'linear' since

$$\begin{aligned} \text{tr}\,\{LM_f^{-1}\} &= \text{tr}\,\{A^T M_f^{-1} A\} \\ &= \text{tr}\,\{\tilde{A}^T M_g^{-1} \tilde{A}\} \\ &= \text{tr}\,\{\tilde{A}\tilde{A}^T M_g^{-1}\} \\ &= \text{tr}\,\{\tilde{L} M_g^{-1}\} \end{aligned}$$

for $L = AA^T$ and $\tilde{L} = \tilde{A}\tilde{A}^T$, with $\tilde{A} = B^{-1}A$.

6.1.2 Linear transformation of the parameters

A D-optimum design is model dependent, however, the design is invariant to non-degenerate linear transformations of the model (Atkinson et al. 2007). This invariance property follows from Theorem 6.1.1 and is proven below.

Theorem 6.1.3 For linear models, a linear transformation of the parameter space is equivalent to a linear transformation of the design space.

Proof For the linear model

$$\mathbb{E}[Y_i] = \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}.$$

Let a linear transformation of the parameters $\boldsymbol{\beta}$ be given by

$$\boldsymbol{\gamma} = B^T\boldsymbol{\beta},$$

with inverse transformation

$$\boldsymbol{\beta} = (B^T)^{-1}\boldsymbol{\gamma} = A^T\boldsymbol{\gamma},$$

where B^T is a $k \times k$ matrix and $|B^T| \neq 0$. The linear model can then be re-expressed as

$$\begin{aligned} \mathbb{E}[Y_i] &= \boldsymbol{f}^T(\boldsymbol{x}_i)A^T\boldsymbol{\gamma} \\ &= \boldsymbol{g}^T(\boldsymbol{x}_i)\boldsymbol{\gamma}, \end{aligned}$$

where

$$\boldsymbol{g}(\boldsymbol{x}_i) = A\boldsymbol{f}(\boldsymbol{x}_i).$$

Clearly, from equation (6.3), this is equivalent to a linear transformation of the design space. □

Corollary 6.1.2 From Theorem 6.1.1, a design D-optimum for the model $\mathbb{E}[Y_i] = \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}$ is also D-optimum for the model $\mathbb{E}[Y_i] = \boldsymbol{g}^T(\boldsymbol{x}_i)\boldsymbol{\gamma}$, if $\boldsymbol{g}(\boldsymbol{x}_i) = A\boldsymbol{f}(\boldsymbol{x}_i)$ and $|A| \neq 0$. However, from Theorem 6.1.2, the trace criterion is not necessarily invariant to linear transformations of the parameter space.

6.2 Admissible Designs with Singular Information Matrices

Recall that the definition of the per observation expected Fisher information matrix is

$$I_i(\boldsymbol{\theta}) = \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right].$$

Using the above information matrix, the information matrix for designing optimum experiments is

$$M(\xi) = \sum w_i I_i(\boldsymbol{\theta}).$$

For the log-linear stochastic frontier model (4.9), and indeed for the more general model (2.2), the first-order partial derivatives of $\ln f_{Y_i} = \ln f_{Y_i}(y_i;\boldsymbol{\theta})$ can be written as

$$\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}_\beta(a_i, \boldsymbol{x}_i) \\ \boldsymbol{f}_\tau(a_i) \end{bmatrix} = \boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i),$$

where a_i is a function of a random variable with realisation ε_i. For example, under a normal-half normal error specification $a_i = \lambda\varepsilon_i/\sigma_G$. The derivative with respect to $\boldsymbol{\beta}$ is written $\boldsymbol{f}_\beta(a_i, \boldsymbol{x}_i)$ because it is a function of both a_i and the explanatory variables $\boldsymbol{x}_i$. For linear models, $\boldsymbol{f}_\beta(a_i, \boldsymbol{x}_i) = f_\beta(a_i)\boldsymbol{f}(\boldsymbol{x}_i)$. Similarly, the derivative with respect to $\boldsymbol{\tau}$ is written $\boldsymbol{f}_\tau(a_i)$ because it is a function of a_i only. Thus the information matrix $M(\xi)$ can be re-expressed as

$$\begin{aligned} M(\xi) &= \sum w_i \mathbb{E}\left[\boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i)\boldsymbol{f}_\theta^T(a_i, \boldsymbol{x}_i)\right] \\ &= \mathbb{E}\left[\sum w_i \boldsymbol{f}_\theta(a_i, \boldsymbol{x}_i)\boldsymbol{f}_\theta^T(a_i, \boldsymbol{x}_i)\right] \\ &= \mathbb{E}\left[\sum w_i \begin{bmatrix} f_\beta(a_i)\boldsymbol{f}(\boldsymbol{x}_i) \\ \boldsymbol{f}_\tau(a_i) \end{bmatrix} \begin{bmatrix} f_\beta(a_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & , & \boldsymbol{f}_\tau^T(a_i) \end{bmatrix}\right] \\ &= \mathbb{E}\begin{bmatrix} \sum w_i f_\beta(a_i)^2 \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \sum w_i \boldsymbol{f}(\boldsymbol{x}_i) f_\beta(a_i)\boldsymbol{f}_\tau^T(a_i) \\ \sum w_i f_\beta(a_i)\boldsymbol{f}_\tau(a_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \sum w_i \boldsymbol{f}_\tau(a_i)\boldsymbol{f}_\tau^T(a_i) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}[f_\beta(a_i)^2]\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\mathbb{E}[f_\beta(a_i)\boldsymbol{f}_\tau^T(a_i)] \\ \mathbb{E}[f_\beta(a_i)\boldsymbol{f}_\tau(a_i)]\sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i) & \mathbb{E}[\boldsymbol{f}_\tau(a_i)\boldsymbol{f}_\tau^T(a_i)] \end{bmatrix}. \end{aligned} \tag{6.4}$$

Note that, like approximated information matrix (6.1), the exact information matrix above is independent of the $\boldsymbol{\beta}$ parameters since expectations are taken over a_i, which is a function of $\varepsilon_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})$. The expectation of a function of a_i is a function of $\boldsymbol{\tau}$ only, that is $\mathbb{E}[f(a_i)] = f(\boldsymbol{\tau})$. Hence optimum designs are independent of $\boldsymbol{\beta}$ but may have a non-simple dependence on $\boldsymbol{\tau}$.

Theorem 6.2.1 For parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$, let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-p})$. If model (2.2) is linear, then its information matrix (6.4) is nonnegative definite with

$$\text{rank } M(\xi) \leq p + 1.$$

Proof Information matrix (6.4) has the same structure as the approximated information matrix (6.1). Consequently, the proof follows the same argument as the proof for Theorem 3.1.2 on approximated information matrices. □

By similar arguments to the proof of Theorem 3.1.2, it is clear from the third equality of equation (6.4) that the last $k - p$ column vectors of the information matrix, associated with the $\boldsymbol{\tau}$ parameters, are not linearly independent, hence the information matrix is singular. Consequently, only subsets or linear combinations of the parameters, say $A^T\boldsymbol{\theta}$, are estimable. Section 5.6 discusses optimum design measures with singular information matrices. Further to Theorem 6.2.1, if the model includes an intercept, β_0, then $p = m + 1$ with

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}) &= [f(x_1), f(x_2), \ldots, f(x_m), 1]^T, \\ \boldsymbol{\beta} &= (\beta_1, \beta_2, \ldots, \beta_m, \beta_0), \end{aligned}$$

and

$$\text{rank } M(\xi) \leq p.$$

If the model does not include β_0 then $p = m$ with

$$\begin{aligned} \boldsymbol{f}(\boldsymbol{x}) &= [f(x_1), f(x_2), \ldots, f(x_m)]^T, \\ \boldsymbol{\beta} &= (\beta_1, \beta_2, \ldots, \beta_m), \end{aligned}$$

and

$$\text{rank } M(\xi) \leq p + 1.$$

In general, rank $M(\xi) \leq m+1$; that is, the rank of the information matrix is equal to the number of $\boldsymbol{\beta}$ parameters, excluding β_0, plus 1. The implication is that it is only possible to design optimally for $(\beta_1, \beta_2, \ldots, \beta_m)$ plus one other parameter, or a linear combination of parameters, from $(\beta_0, \boldsymbol{\tau})$, or in general, a linear combination $A^T\boldsymbol{\theta}$. A matrix A which ensures a nonsingular matrix $A^T M^- A$ has the form

$$A = \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{a} \end{bmatrix}, \tag{6.5}$$

where A_{11} is a $m \times m$ matrix with $|A_{11}| \neq 0$ and $\boldsymbol{a}$ is a column vector of length $(k-p+1)$. Let $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ and $\tilde{\boldsymbol{\tau}} = (\beta_0, \boldsymbol{\tau})$ so that $\boldsymbol{\theta} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})$, then the set of admissible linear combinations of parameters is

$$A^T\boldsymbol{\theta} = \begin{bmatrix} A_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{a} \end{bmatrix}^T \begin{bmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\boldsymbol{\tau}} \end{bmatrix} = \begin{bmatrix} A_{11}^T\tilde{\boldsymbol{\beta}} \\ \boldsymbol{a}^T\tilde{\boldsymbol{\tau}} \end{bmatrix}. \tag{6.6}$$

6.3 Equivalence of Transformations

The model for a log-linear stochastic production frontier can be expressed as

$$\begin{aligned} \mathbb{E}[Y_i] &= \beta_0^{**} + \sum_{j=1}^{m} \beta_j x_{ij} \\ &= \boldsymbol{f}^T(\boldsymbol{x}_i) B^T \boldsymbol{\beta}, \end{aligned} \tag{6.7}$$

where $\boldsymbol{f}(\boldsymbol{x}_i) = (x_{i1}, x_{i2}, \ldots, x_{im}, 1)^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m, \beta_0)^T$. When an intercept term, β_0, is in the frontier model, $\beta_0^{**} = \beta_0 - \mathbb{E}[U_i]$ and

$$B = \begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & 1 - \mathbb{E}[U_i]/\beta_0 \end{bmatrix},$$

where I_m is the $m \times m$ identity matrix. When β_0 is not in the frontier model,

$\beta_0^{**} = -\mathbb{E}[U_i]$ and

$$B = \begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & -\mathbb{E}[U_i]/\beta_0 \end{bmatrix}.$$

From the first equality in equation (6.7), a log-linear stochastic production frontier model can be viewed as a regression model with a shifted intercept β_0^{**}. The effect of the shifted intercept can be seen in the second equality which demonstrates that the frontier model is a regression model with a transformation applied to the β_0 parameter.

Model (6.7) can be re-expressed, in terms of the parameter vector $\boldsymbol{\theta} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\tau}})$, as

$$\mathbb{E}[Y_i] = \boldsymbol{f}^T(\boldsymbol{x}_i) A^T \boldsymbol{\theta},$$

where $\boldsymbol{f}(\boldsymbol{x}_i) = (x_{i1}, x_{i2}, \ldots, x_{im}, 1)^T$ is unchanged and A is the $k \times (m+1)$ matrix (6.5) of rank $m+1 < k$. The matrix A can be selected to give $A^T\boldsymbol{\theta} = B^T\boldsymbol{\beta}$.

6.3.1 Normal-half normal model

A log-linear stochastic production frontier model with a normal-half normal error specification has

$$\tilde{\boldsymbol{\tau}} = \begin{bmatrix} \beta_0, & \lambda, & \sigma_G^2 \end{bmatrix}^T,$$

where $\lambda = \sigma_u/\sigma_v$ and $\sigma_G^2 = \sigma_u^2 + \sigma_v^2$. The expected value of U is

$$\mathbb{E}[U] = \sqrt{\frac{2}{\pi}}\sigma_u = \sqrt{\frac{2}{\pi}}\frac{\lambda\sigma_G}{(\lambda^2+1)^{1/2}}.$$

For matrix A given in equation (6.5), $A_{11} = I_m$ and either

$$\boldsymbol{a} = \begin{bmatrix} 1, & -\sqrt{\frac{2}{\pi}}\frac{\sigma_G}{(\lambda^2+1)^{1/2}}, & 0 \end{bmatrix}^T, \tag{6.8}$$

or

$$\boldsymbol{a} \;=\; \left[\; 1, \;\; 0, \;\; -\sqrt{\frac{2}{\pi}}\frac{\lambda}{\sigma_G(\lambda^2+1)^{1/2}} \;\right]^T, \tag{6.9}$$

will give $A^T\boldsymbol{\theta} = B^T\boldsymbol{\beta}$. The linear combination $\boldsymbol{a}^T\tilde{\boldsymbol{\tau}}$ is the shifted intercept $\beta_0^{**} = \beta_0 - \mathbb{E}[U]$.

6.3.2 Normal-exponential model

A normal-exponentially distributed frontier model has

$$\tilde{\boldsymbol{\tau}} \;=\; \left[\; \beta_0, \;\; 1/\sigma_u, \;\; \sigma_v^2 \;\right]^T,$$

and expected value of U given by

$$\mathbb{E}[U] \;=\; \sigma_u.$$

For matrix A given in equation (6.5), $A_{11} = I_m$ and

$$\boldsymbol{a} \;=\; \left[\; 1, \;\; -\sigma_u^2, \;\; 0 \;\right]^T, \tag{6.10}$$

will give $A^T\boldsymbol{\theta} = B^T\boldsymbol{\beta}$.

6.4 Optimum Designs using Determinant Criterion Functions

For polynomial regression in one variable, Atkinson et al. (2007) give the points of support of D-optimum designs for m-th order polynomials

$$\mathbb{E}[Y] \;=\; \beta_0 + \sum_{j=1}^{m}\beta_j x^j,$$

for $m = 2, \ldots, 6$. For $\mathscr{X} = [-1, 1]$, when $m = 1$, the optimum design places half the trials at $x = 1$ and the other half at $x = -1$. When $m = 2$, the optimum

design places a third of the trials at $x = -1$, 0 and 1. In general, for $p = m+1$, the design puts mass $1/p$ at p distinct design points. Happacher (1995) reports some results for exact and continuous D-optimum designs for polynomial regression with degree ≤ 40.

The D-optimum design in not invariant to removal of the intercept term, β_0, from the polynomial regression model. Chang (1999) reports some results on D-optimum designs over design space $\mathscr{X} = [a, 1]$, $-1 \leq a < 1$, for polynomial regression in one variable through the origin. The following example illustrates the differences in optimum designs between a polynomial regression model with and without an intercept.

Example 6.4.1 Quadratic regression in one variable.

The D-optimum design over design space $\mathscr{X} = [a, 1]$, $-1 \leq a < 1$, for model $\boldsymbol{f}(x) = (x, x^2, 1)^T$ is

$$\xi^* = \left\{ \begin{array}{ccc} a & \frac{1+a}{2} & 1 \\ 1/3 & 1/3 & 1/3 \end{array} \right\}.$$

This design puts equal mass at the three equally spaced support points where the lower and upper support points are at the boundary of the design space. Removal of the intercept term from the model produces different D-optimum designs. For model $\boldsymbol{f}(x) = (x, x^2)^T$, if $n \geq 2$ and $(2 - \sqrt{10})/6 \leq a \leq 1/2$, Chang (1999) reports that the optimum design is

$$\xi^* = \left\{ \begin{array}{cc} 1/2 & 1 \\ 1/2 & 1/2 \end{array} \right\}.$$

Note that this design puts equal weights at two support points but that the lower interior support point is not at the boundary of the design region, unless $a = 1/2$. If $1/2 \leq a < 1$, the optimum design is

$$\xi^* = \left\{ \begin{array}{cc} a & 1 \\ 1/2 & 1/2 \end{array} \right\},$$

that is, the design puts equal weights at two support points where the support points are at the boundary of the design space. Removal of the intercept term from the regression model reduces the number of parameters, hence potentially the number of support points, by one. Additionally, the support points for the reduced model may not be at the boundary of the design region. □

Unlike regression models, removal of the intercept term from a stochastic frontier model does not change the optimum design under a determinant criterion. Theorem 6.1.1 gives a powerful result on D_A-optimum designs for log-linear stochastic production frontier models, which is stated in the following corollary.

6.4.1 Equivalence of designs for regression and frontier models

Corollary 6.4.1 [to Theorem 6.1.1] A design D_A-optimum for a log-linear stochastic production frontier model, with or without an intercept term, is also D_A-optimum for the corresponding regression model $\mathbb{E}[Y_i] = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}$ that has an intercept term.

Proof Section 6.3 demonstrated that the frontier model is a linear regression model with a shifted intercept. Thus it might be expected that the D_A-optimum design will be that for a regression model with an intercept. Additionally, the frontier model is a regression model with a linear transformation applied to β_0. Consequently, Theorem 6.1.1 gives the results that both models are maximised by the same D_A-optimum design, due to the invariance property of D-optimum designs to linear transformations of the parameters. □

The corollary amounts to stating that, if an experimenter wishes to design a D_A-optimum design for a log-linear stochastic production frontier model, it

will be the same design as a D_A-optimum design for the corresponding regression model with an intercept. This equivalence implies that the D_A-optimum design for optimal estimation of $A_{11}^T\tilde{\boldsymbol{\beta}}$ and *any* linear combination $\boldsymbol{a}^T\tilde{\boldsymbol{\tau}}$ from equation (6.6), is the D_A-optimum design for optimal estimation of $A_{11}^T\tilde{\boldsymbol{\beta}}$ and the shifted intercept β_0^{**}. Since D_A-optimum designs for linear regression models are independent of the parameters, so too are D_A-optimum designs for log-linear stochastic production frontier models. Another implication of the corollary is that D_A-optimum designs for the frontier model are independent of any distributional assumption imposed on the efficiency term U.

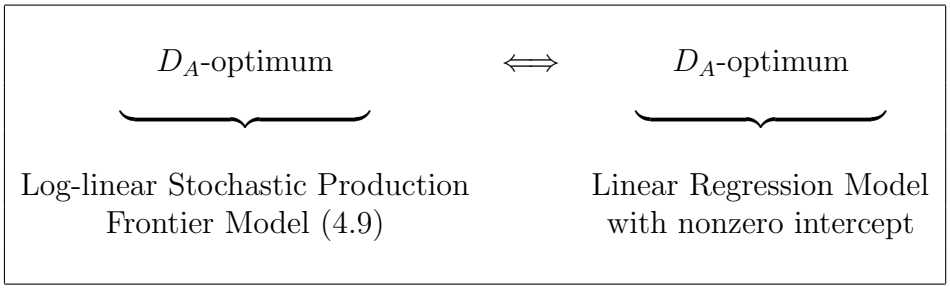

Figure 6.1: Equivalence of D_A-optimum designs for log-linear stochastic production frontier models and linear regression models with nonzero intercept.

Example 6.4.2 Quadratic regression in one variable.

The D_A-optimum design for the parameters $(\beta_1, \beta_2, \beta_0^{**})$ has

$$A = \begin{bmatrix} I_m & \mathbf{0} \\ \mathbf{0} & \boldsymbol{a} \end{bmatrix},$$

where $\boldsymbol{a}$ is given in equations (6.8) and (6.9) for a normal-half normal frontier model, and in equation (6.10) for a normal-exponential model. For $\mathscr{X} = [0, 1]$ the optimum design is

$$\xi^* = \begin{Bmatrix} 0 & 1/2 & 1 \\ 1/3 & 1/3 & 1/3 \end{Bmatrix}.$$

From Example 6.4.1, this is the D_A-optimum design for $(\beta_1, \beta_2, \beta_0)$ in a linear regression model with an intercept. Note that prior values of σ_u and σ_v are required to calculate the elements of $\boldsymbol{a}$. However, by Corollary 6.4.1, the D_A-optimum design for the linear regression model is D_A-optimum for the equivalent frontier model, hence the elements of the vector $\boldsymbol{a}$ can take any values for the determinant criterion.

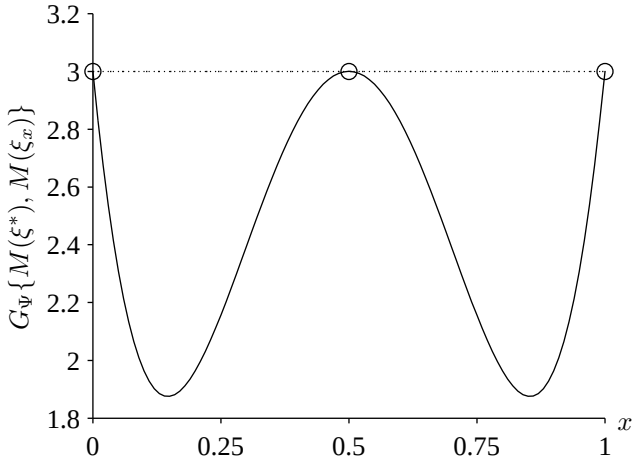

Figure 6.2: Example 6.4.2: quadratic regression in one variable. Gâteaux derivative $G_\Psi\{M(\xi^*), M(\xi_x)\}$ for the D_A-optimum design for $(\beta_1, \beta_2, \beta_0^{**})$ where $G_\Psi\{M(\xi^*), M(\xi_i)\} = 3$.

Figure 6.2 confirms that the design is optimal by the General Equivalence Theorem, since $G_\Psi\{M(\xi^*), M(\xi)\} \leq 3$ for all $\xi \in \Xi$ and achieves its maximum at the points of support of the design. □

The results on D-optimum designs for regression models are well established, hence will not be given in any further detail here.

6.5 Optimum Designs using Trace Criterion Functions

Pukelsheim (1980) and Pukelsheim & Torsney (1991) give the A-optimum de-

sign for $(\beta_1, \beta_2, \beta_0)$ and the C-optimum design for (β_1, β_2) for quadratic regression in one variable over the symmetric interval $\mathscr{X} = [-1, 1]$. The optimum designs are presented in the following example.

Example 6.5.1 Quadratic regression in one variable.

For model $\boldsymbol{f}(x) = (x, x^2, 1)$, ξ_3^* gives the A-optimum design for $(\beta_1, \beta_2, \beta_0)$ and ξ_2^* gives the C-optimum design for (β_1, β_2) over $\mathscr{X} = [-1, 1]$.

$$\xi_3^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 1/4 & 1/2 & 1/4 \end{array} \right\} \qquad \xi_2^* = \left\{ \begin{array}{ccc} -1 & 0 & 1 \\ 0.2929 & 0.4142 & 0.2929 \end{array} \right\}$$

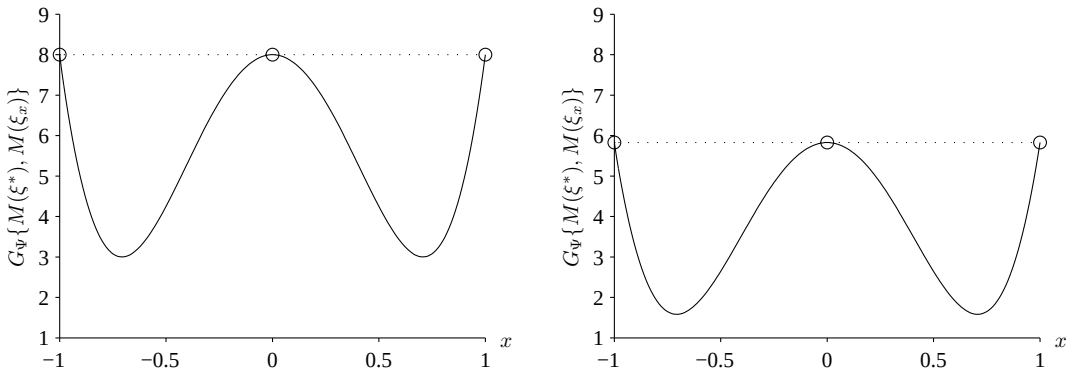

Figure 6.3: Example 6.5.1: quadratic regression in one variable. Gâteaux derivative $G_\Psi\{M(\xi^*), M(\xi_x)\}$ for the C-optimum design over $\mathscr{X} = [-1, 1]$ for; (a) $(\beta_1, \beta_2, \beta_0)$ where $G_\Psi\{M(\xi^*), M(\xi_i)\} = 8$; (b) (β_1, β_2) where $G_\Psi\{M(\xi^*), M(\xi_i)\} = 5.83$

Figures 6.3 (a) and (b) confirm that the respective designs ξ_3^* and ξ_2^* are optimal by the General Equivalence Theorem. Both designs are symmetric in the support points and the weights, although the weights differ between the two designs. □

Unlike the determinant criterion, the trace criterion is not invariant to linear transformations. Hence when the design region is not symmetric, the weights

may no longer be symmetric. This can be seen in the following example where the design interval is asymmetric.

Example 6.5.2 Quadratic regression in one variable (Example 6.5.1 continued).

This example follows Example 6.5.1 but here the designs are over the asymmetric interval $\mathscr{X} = [0, 1]$.

$$\xi_3^* = \left\{ \begin{array}{ccc} 0 & 1/2 & 1 \\ 0.3216 & 0.4862 & 0.1922 \end{array} \right\} \qquad \xi_2^* = \left\{ \begin{array}{ccc} 0 & 1/2 & 1 \\ 0.3136 & 0.4920 & 0.1944 \end{array} \right\}$$

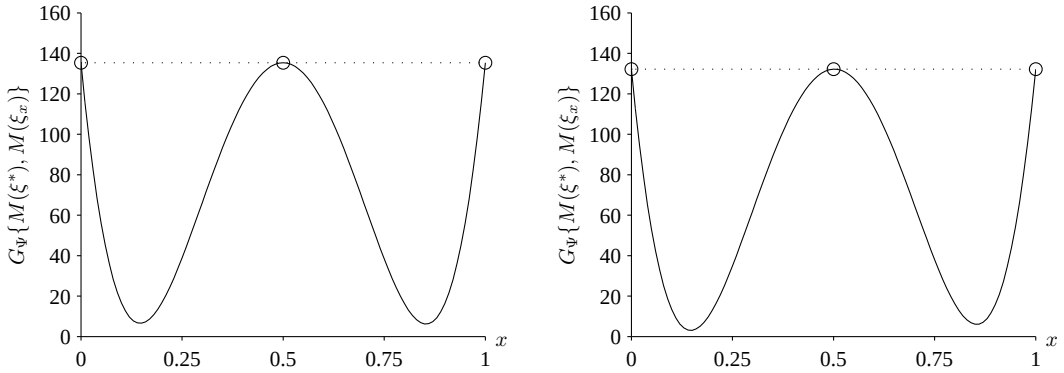

Figure 6.4: Example 6.5.2: quadratic regression in one variable. Gâteaux derivative $G_\Psi\{M(\xi^*), M(\xi_x)\}$ for the C-optimum design over $\mathscr{X} = [0, 1]$ for; (a) $(\beta_1, \beta_2, \beta_0)$ where $G_\Psi\{M(\xi^*), M(\xi_i)\} = 135.36$; (b) (β_1, β_2) where $G_\Psi\{M(\xi^*), M(\xi_i)\} = 132.21$

Figures 6.4 (a) and (b) confirm that the respective designs ξ_3^* and ξ_2^* are optimal by the General Equivalence Theorem. Both designs are symmetric in the support points, however the weights are no longer symmetric. □

6.5.1 Linear C-optimum designs for the β parameters, excluding β_0

From Section 6.3, a log-linear stochastic frontier model can be viewed as a regression model with a transformation applied to the β_0 parameters. Hence if interest is in estimating all the $\boldsymbol{\beta}$ parameters, *excluding* β_0, then the C-optimum design for $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ for a regression model is also C-optimum for a log-linear stochastic production frontier model. That is, the C-optimum design will be that for a regression model but the value of the criterion function will differ. The criterion function for the stochastic frontier model is likely to increase with increasing values of σ_u and σ_v. However, the values of σ_u and σ_v are irrelevant in finding the optimum design. Since the C-optimum design for the linear regression model is independent of the parameters, so too is the C-optimum design for the log-linear stochastic production frontier model. The designs are also independent of any distributional assumption imposed on the efficiency term U.

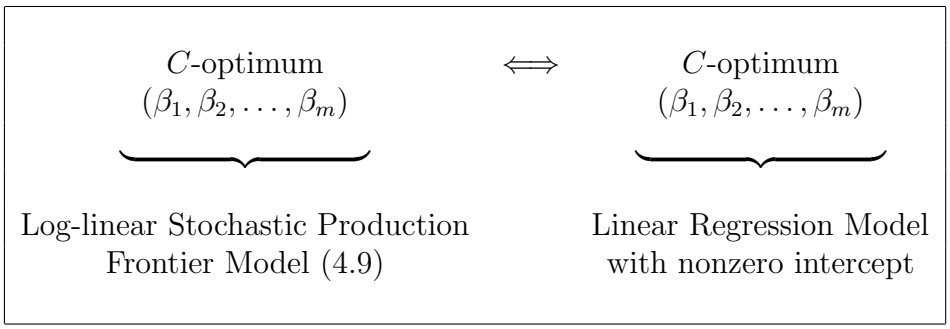

Figure 6.5: Equivalence of C-optimum designs for $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ for log-linear stochastic production frontier models and linear regression models with nonzero intercept.

Examples 6.5.1 and 6.5.2 for the quadratic regression model thus give some results on C-optimum designs for log-linear stochastic production frontier models. For a quadratic stochastic frontier model in one variable, the C-optimum design over the symmetric interval $\mathscr{X} = [-1, 1]$ is ξ_2^* from Example 6.5.1 and the C-

optimum design over the asymmetric interval $\mathscr{X} = [0, 1]$ is ξ_2^* from Example 6.5.2.

6.5.2 Nonlinear C-optimum designs for the β parameters and shifted intercept β_0^{**}

If interest is in estimating the $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ parameters *and* the shifted intercept β_0^{**}, then standard results from designs for linear regression do not carry over. This is because the trace criterion is not invariant to linear transformations, in this case, of the β_0 parameter. Hence the C-optimum design for $(\beta_1, \beta_2, \ldots, \beta_m, \beta_0^{**})$ is dependent on (i) the design region, (ii) the distributional assumption imposed on the efficiency term U, (iii) the matrix $C(= A)$ used to define the contrast $A^T\boldsymbol{\theta}$ and (iv) the values of the parameters σ_u and σ_v. The dependence of the C-optimum design on these four factors is demonstrated in the following examples where Torsney's (1977) algorithm (5.13) was implemented to find the optimising weights over a 26×26 grid of σ_u and σ_v values with $0.1 \leq \sigma_u, \sigma_v \leq 0.35$ in increments of 0.01.

Example 6.5.3 Quadratic regression in one variable.

For model $\boldsymbol{f}(x) = (x, x^2, 1)$, C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over the symmetric interval $\mathscr{X} = [-1, 1]$ are presented in Figures 6.6, 6.7, and 6.8; Figure 6.6 gives the optimum designs under a normal-half normal error specification with $\boldsymbol{a}$ given by equation (6.9); Figure 6.7 gives the optimum designs under a normal-half normal error specification with $\boldsymbol{a}$ given by equation (6.8); Figure 6.8 gives the optimum designs under a normal-exponential error specification with $\boldsymbol{a}$ given by equation (6.10).

Subplot (a) demonstrates that the optimum designs are supported on three symmetric and evenly spaced points at -1, 0 and 1 for all designs. Subplots (b), (c) and (d) give the surface of the design weights w_1^*, w_2^* and w_3^*, respectively, over the 26×26 grid of σ_u and σ_v values. In each subplot, the surface is not

a flat plane, hence the C-optimum design is different for varying values of the parameters σ_u and σ_v. Comparing subplots (b), (c) and (d) across (i) Figures 6.6 and 6.8, and (ii) Figures 6.7 and 6.8, shows that the surface for each weight differs depending on the distributional assumption placed on the efficiency term U. A comparison of subplots (b), (c) and (d) across Figures 6.6 and 6.7 shows that the surface for each weight differs depending on the matrix A used in the contrast $A^T\boldsymbol{\theta}$ for a normal-half normal error specification. Subplots (b) and (d) for optimal design weights w_1^* and w_3^*, respectively, depict the same surface indicating that the design weights are symmetric.

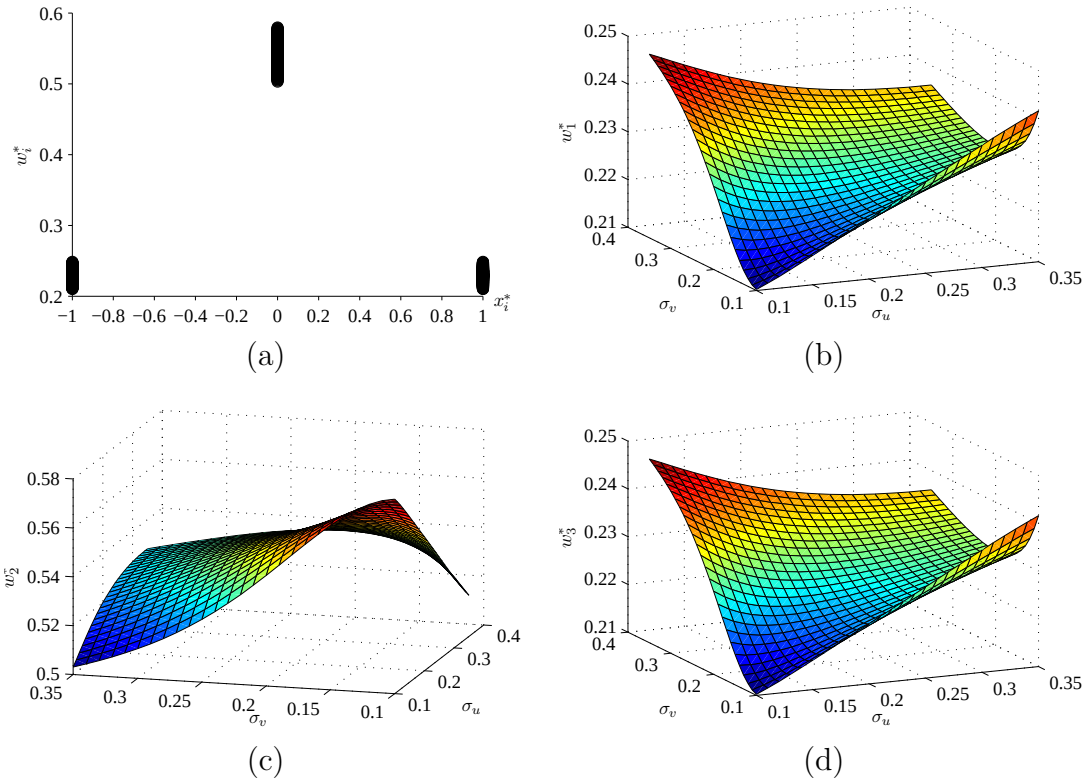

Figure 6.6: Example 6.5.3: quadratic regression in one variable. C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over $\mathscr{X} = [-1, 1]$ under a normal-half normal error specification with $\boldsymbol{a} = [1, 0, -\sqrt{2/\pi}\lambda/(\sigma_G(\lambda^2+1)^{1/2})]^T$ and $0.1 \leq \sigma_u, \sigma_v \leq 0.35$; (a) optimal weights vs. optimal support points; (b) distribution of w_1^*; (c) distribution of w_2^*; (d) distribution of w_3^*.

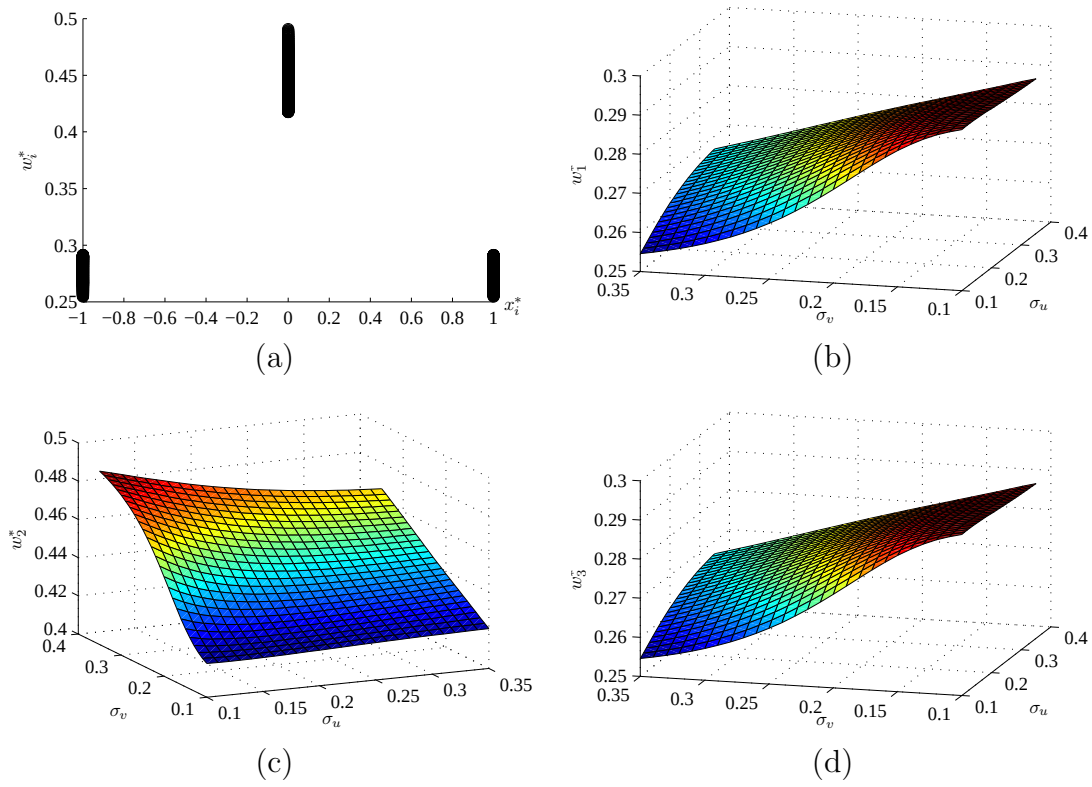

Figure 6.7: Example 6.5.3: quadratic regression in one variable. C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over $\mathscr{X} = [-1, 1]$ under a normal-half normal error specification with $\boldsymbol{a} = [1, -\sqrt{2/\pi}\sigma_G/(\lambda^2+1)^{1/2}), 0]^T$ and $0.1 \leq \sigma_u, \sigma_v \leq 0.35$; (a) optimal weights vs. optimal support points; (b) distribution of w_1^*; (c) distribution of w_2^*; (d) distribution of w_3^*.

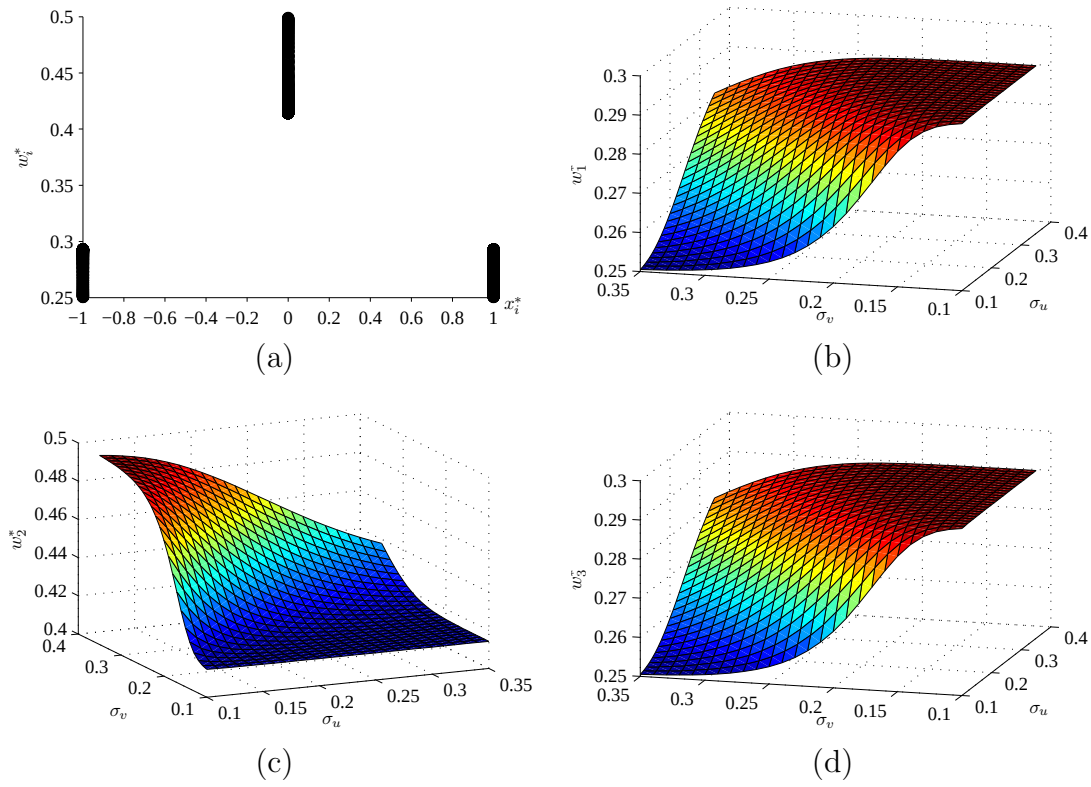

Figure 6.8: Example 6.5.3: quadratic regression in one variable. C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over $\mathscr{X} = [-1, 1]$ under a normal-exponential error specification with $\boldsymbol{a} = [1, -\sigma_u^2, 0]^T$ and $0.1 \leq \sigma_u, \sigma_v \leq 0.35$; (a) optimal weights vs. optimal support points; (b) distribution of w_1^*; (c) distribution of w_2^*; (d) distribution of w_3^*.

□

Example 6.5.4 Quadratic regression in one variable (Example 6.5.3 continued).

This example follows Example 6.5.3 but here the designs are over the asymmetric interval $\mathscr{X} = [0, 1]$. Figure 6.9 gives the optimum designs under a normal-half normal error specification with $\boldsymbol{a}$ given by equation (6.9); Figure 6.10 gives the optimum designs under a normal-half normal error specification with $\boldsymbol{a}$ given by equation (6.8); Figure 6.11 gives the optimum designs under a normal-exponential error specification with $\boldsymbol{a}$ given by equation (6.10).

As with the symmetric design region of Example 6.5.3, subplot (a) demonstrates that the optimum designs are supported on three symmetric and evenly spaced points. Here the support points are at 0, 1/2 and 1 for all designs. Comparisons that were made in Example 6.5.3 can be made here. A summary of the comparisons is similar to that given for the previous example; (i) the surface for each weight is not a flat plane, hence the C-optimum design is different for varying values of the parameters σ_u and σ_v; (ii) the surface for each weight differs depending on the distributional assumption placed on the efficiency term U; (iii) the surface for each weight differs depending on the matrix A used in the contrast $A^T\boldsymbol{\theta}$ for a normal-half normal error specification.

However, subplots (b) and (d) for optimal design weights w_1^* and w_3^*, respectively, do *not* depict a common surface, as they did in Example 6.5.3, indicating that the design weights are asymmetric on the asymmetric interval $\mathscr{X} = [0, 1]$. Further, comparing (i) Figures 6.6 and 6.9, (ii) Figures 6.7 and 6.10, and (iii) Figures 6.8 and 6.11, gives a comparison of designs over design region $\mathscr{X} = [-1, 1]$ and $\mathscr{X} = [0, 1]$. Such a comparison shows that the surface of each design weight differs depending on the design region. Hence the C-optimum design is not invariant to translational changes in the design region $\mathscr{X}$.

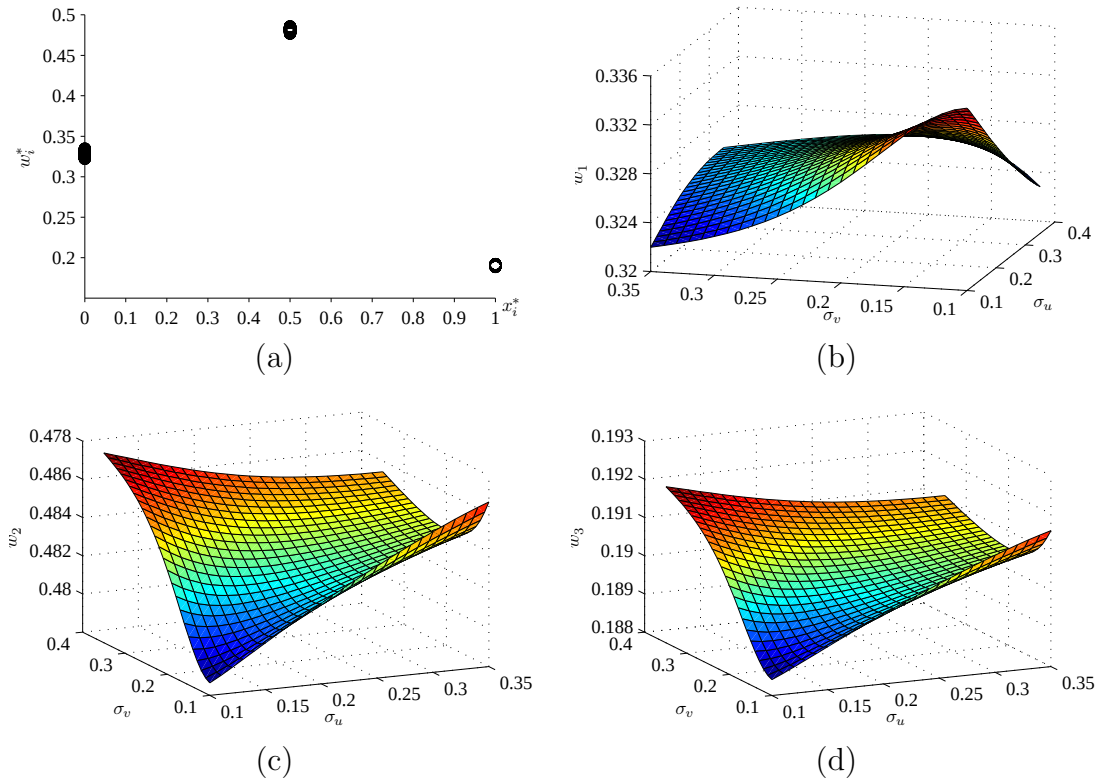

Figure 6.9: Example 6.5.4: quadratic regression in one variable. C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over $\mathscr{X} = [0, 1]$ under a normal-half normal error specification with $\boldsymbol{a} = [1, 0, -\sqrt{2/\pi}\lambda/(\sigma_G(\lambda^2+1)^{1/2})]^T$ and $0.1 \leq \sigma_u, \sigma_v \leq 0.35$; (a) optimal weights vs. optimal support points; (b) distribution of w_1^*; (c) distribution of w_2^*; (d) distribution of w_3^*.

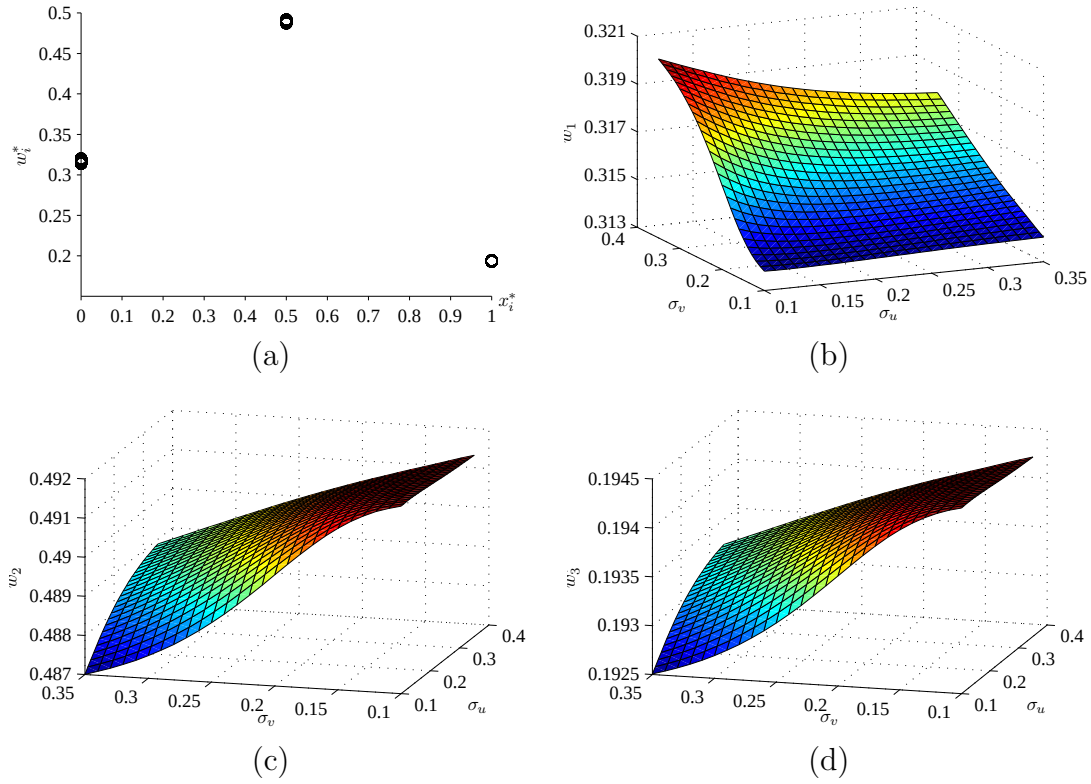

Figure 6.10: Example 6.5.4: quadratic regression in one variable. C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over $\mathscr{X} = [0, 1]$ under a normal-half normal error specification with $\boldsymbol{a} = [1, -\sqrt{2/\pi}\sigma_G/(\lambda^2+1)^{1/2}), 0]^T$ and $0.1 \leq \sigma_u, \sigma_v \leq 0.35$; (a) optimal weights vs. optimal support points; (b) distribution of w_1^*; (c) distribution of w_2^*; (d) distribution of w_3^*.

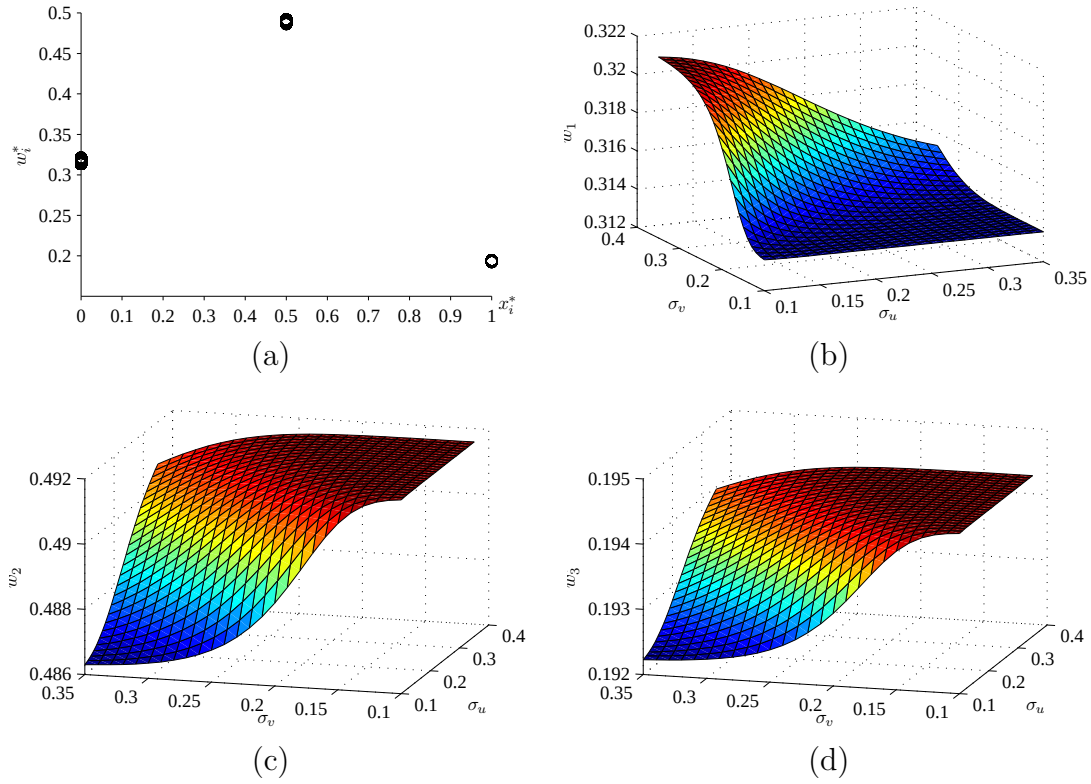

Figure 6.11: Example 6.5.4: quadratic regression in one variable. C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$ over $\mathscr{X} = [0, 1]$ under a normal-exponential error specification with $\boldsymbol{a} = [1, -\sigma_u^2, 0]^T$ and $0.1 \leq \sigma_u, \sigma_v \leq 0.35$; (a) optimal weights vs. optimal support points; (b) distribution of w_1^*; (c) distribution of w_2^*; (d) distribution of w_3^*.

□

Designs with approximated and exact information matrices

The designs calculated in Examples 6.5.3 and 6.5.4 were found using the approximated information matrix (6.1). The C-optimum design depends on the parameters σ_u and σ_v which appear inside functions (or expectations of functions) in each block of the partitioned information matrix. These functions are different for the approximated information matrix (6.1) and the exact information matrix (6.4). Hence the optimum design is also likely to depend on the approximation method implemented.

Remark Some theoretical results on optimum designs are not affected by the approximation method applied to the information matrix since the structure of the approximated information matrix (6.1) and the exact information matrix (6.4) are similar. Where differences arise, they are due to differences between the approximations in how they depend on the parameters σ_u and σ_v. For example, any results based on the rank of the information matrix does not depend on the approximation method used.

Chapter 7

Conclusions

7.1 A Model with Skewed Composed Error

Optimum designs for a general, possibly nonlinear, statistical model

$$Y = f(\boldsymbol{x}, \boldsymbol{\beta}) + E, \qquad \mathbb{E}[E] \neq 0, \tag{7.1}$$

with skewed asymmetrically distributed error E, have been explored and designs developed for the linear case of this model, written

$$Y = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} + E, \qquad \mathbb{E}[E] \neq 0. \tag{7.2}$$

Since random error, say V, is typically assumed to be symmetrically distributed, then if an overall error E is to be asymmetrically distributed, it can be viewed as a symmetric random error V and some other source of error, say U, that is asymmetrically distributed. That is, the overall or 'composed' error can be modelled as a linear combination of a symmetric random error term V and an asymmetric error term U, written

$$E = c_u U + c_v V, \qquad \mathbb{E}[U] \neq 0,\ \mathbb{E}[V] = 0,\ \{c_u, c_v\} \in \mathbb{R}.$$

Linear models with this type of error structure appear in the econometric literature where V is normally distributed with zero mean and common distributions

for the error term U are the nonnegative half normal, exponential, nonnegative truncated normal and gamma distributions. Hence the common distributions of E are called normal-half normal, normal-exponential, normal-truncated normal and normal-gamma, respectively.

7.2 Derivation of the Information Matrix

The per observation expected Fisher information matrix is required for the design of optimum experiments. The information matrix of the full parameter vector $\boldsymbol{\theta}$ was derived in Chapter 2 for general model (7.1) under the four error specifications mentioned above. The information matrix for linear model (7.2) can be easily found by letting $f(\boldsymbol{x}, \boldsymbol{\beta}) = \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta}$ in the information matrix. The process of deriving the information matrix is as follows: obtain the joint density of U and V; by a transformation of variables, obtain the joint density of U and E; integrate this joint density with respect to u to obtain the marginal density, $f(\varepsilon)$, of E, and its mean and variance; calculate the per observation log-likelihood function, $\ln f(y; \boldsymbol{\theta})$, of the full parameter vector $\boldsymbol{\theta}$ using $f(\varepsilon)$; calculate the first-order and second-order partial derivatives of $\ln f(y; \boldsymbol{\theta})$ with respect to the parameters; use these partial derivatives to obtain the per observation expected Fisher information matrix using definition (D.2) or (D.3) in Appendix D.2. Additionally, the conditional density of $U|E$, and its mean and mode were also derived. They are not required in the derivation of the information matrix but are required to obtain measures of efficiency.

7.3 Structure of the Information Matrix

7.3.1 Nonlinearity of designs with parameter dependent information matrices

Although model (7.2) is *linear*, inspection of the information matrix reveals that optimum designs for this model may be *parameter dependent*, due to the asymmetric distribution of the composed error. This is an interesting feature of this model since parameter dependent nonlinear optimum designs typically arise from nonlinear models. The elements of information matrix (6.4) of the extended parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ for model (7.2) are functions of the $\boldsymbol{\tau}$ parameters, but not the $\boldsymbol{\beta}$ parameters, hence the optimum design may depend on the $\boldsymbol{\tau}$ parameters but does not depend on the $\boldsymbol{\beta}$ parameters.

7.3.2 Admissible designs with singular information matrices

The $k \times k$ information matrix M, derived in the manner discussed above, is rank deficient. A consequence of the singularity of the information matrix is that only subsets or s linear combinations of the parameters, $A^T\boldsymbol{\theta}$, where A is a matrix of rank $s < k$ which makes $A^T M^- A$ nonsingular, are estimable. The matrix inverse M^- can be *any* generalised inverse, however, the Moore-Penrose generalised inverse, also called the pseudoinverse or the (1,2,3,4)-inverse, was used since it is a unique generalised inverse. The rank of the partitioned information matrix is such that admissible designs are only possible for optimal estimation of

$$A^T\boldsymbol{\theta} \;=\; \begin{bmatrix} A_{11}^T\tilde{\boldsymbol{\beta}} \\ \boldsymbol{a}^T\tilde{\boldsymbol{\tau}} \end{bmatrix},$$

that is, at best, the parameters $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ and one other parameter, or linear combination of parameters, from $\tilde{\boldsymbol{\tau}} = (\beta_0, \boldsymbol{\tau})$. The asymmetrically

distributed linear model (7.2), which can be written as

$$\mathbb{E}[Y] \;=\; \beta_0^{**} + \sum_{j=1}^{m} \beta_j x_j, \tag{7.3}$$

is equivalent to the usual symmetrically distributed linear regression model

$$\mathbb{E}[Y] \;=\; \beta_0 + \sum_{j=1}^{m} \beta_j x_j, \tag{7.4}$$

but with a shifted intercept $\beta_0^{**} = \beta_0 - \mathbb{E}[U]$. The shifted intercept is equivalent to a transformation applied to the parameter β_0 in the usual linear regression model. This transformation, which depends on the distributional assumption on U, provides guidance on appropriate choices for the vector $\boldsymbol{a}$ to give linear combinations $\boldsymbol{a}^T\tilde{\boldsymbol{\tau}} = \beta_0^{**}$. Thus appropriate and admissible linear combinations are given by

$$A^T\boldsymbol{\theta} \;=\; \begin{bmatrix} A_{11}^T\tilde{\boldsymbol{\beta}} \\ \beta_0^{**} \end{bmatrix}.$$

Any other choice of linear combination $\boldsymbol{a}^T\tilde{\boldsymbol{\tau}} \neq \beta_0^{**}$, although feasible, gives a biased estimate of the shifted intercept β_0^{**}.

7.4 Linear D_A- and C-Optimum Designs

Since D_A-optimum designs are invariant to linear transformations, designs that maximise the criterion function for D_A-optimality for the usual linear regression model (7.4) with nonzero intercept also maximise the D_A-criterion function for asymmetrically distributed linear model (7.3). Trace criterion functions are not invariant to linear transformations, in this case, of the parameter β_0. However, if interest is in estimating the $\boldsymbol{\beta}$ parameters, *excluding* β_0, then the C-optimum design criterion for $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ for the asymmetrically distributed model (7.3) is maximised by the C-optimum design for $\tilde{\boldsymbol{\beta}}$ for the usual

linear regression model (7.4) with nonzero intercept. Hence the D_A-optimum design, and C-optimum design for $\tilde{\boldsymbol{\beta}}$, for the asymmetrically distributed linear model are, respectively, just the standard linear, non-parameter dependent, D_A-optimum design, and C-optimum design for $\tilde{\boldsymbol{\beta}}$, for the usual symmetrically distributed linear regression model with nonzero intercept.

7.5 Nonlinear C-Optimum Designs

If interest is in designing an experiment for optimal estimation of $\tilde{\boldsymbol{\beta}} = (\beta_1, \beta_2, \ldots, \beta_m)$ *and* the shifted intercept β_0^{**}, then C-optimum designs for $(\beta_1, \beta_2, \ldots, \beta_m, \beta_0^{**})$ are not standard linear C-optimum designs for the usual symmetrically distributed linear regression model. The nonlinear C-optimum design depends on: (i) the design region $\mathscr{X}$; (ii) the distributional assumption on the efficiency term U; (iii) the matrix A used to define the contrast $A^T\boldsymbol{\theta}$ for admissible designs; (iv) the variance parameters σ_u and σ_v; (v) the method used to approximate the information matrix.

7.6 Special Case: Log-Linear Cobb-Douglas Stochastic Production Frontier Model

A special case of model (7.2) is the econometric cross-sectional Cobb-Douglas stochastic frontier model

$$\ln Y \;=\; \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_j + E.$$

This is just the asymmetrically distributed linear regression model (7.3) with a logarithmic transformation applied to the observed response y and predictors x_j. Stochastic frontier models are used to obtain relative measures of efficiency for organisations with similar characteristics, e.g. within the same industry. Various

measurements of efficiency that were discussed in Chapter 4 included: input-oriented and output oriented efficiency; technical and economic efficiency; efficiency measured relative to a 'frontier'; and parametric and nonparametric measures of efficiency. These measurements are not necessarily mutually exclusive but are various ways of classifying efficiency.

The type of model used in the examples of Chapter 6 is a single-output cross-sectional stochastic production frontier models used to obtain measures of output-oriented technical efficiency. For this model the composed error has the structure $E = V - U$, giving the model

$$\ln Y \quad = \quad \beta_0 + \sum_{j=1}^{m} \beta_j \ln x_j + V - U.$$

Of the four common distributions of $E = V - U$, the simpler normal-half normal and normal-exponential distributions are preferable (Ritter & Simar 1997). Since this model is a special case of the more general model (7.1), the formulae of Chapter 2 can be easily simplified to give the appropriate information matrix for the stochastic production frontier model. These simplifications were carried out in Chapter 4 to give the corresponding information matrices, which can be used to obtain optimum designs for the frontier model. Formulae for calculating the average efficiency across an industry, and the efficiency of individual organisations relative to each other, were also derived for completeness.

In Chapter 6, some numerical results for a quadratic frontier model in one variable were presented to demonstrate the nonlinearity of C-optimum designs for $(\beta_1, \beta_2, \beta_0^{**})$. The numerical results also demonstrate the dependence of the optimum design on the design region $\mathscr{X}$, the distributional assumption on the efficiency term U, and the matrix A used to define the contrast $A^T\boldsymbol{\theta}$ for admissible designs. The parameters σ_u and σ_v were allowed to vary over the equally spaced 26×26 grid on [0.1,0.35]. Torsney's (1977) algorithm (5.13) was implemented to find the optimising design weights. The C-optimum designs over symmetric

design region $\mathscr{X} = [-1, 1]$ and asymmetric design region $\mathscr{X} = [0, 1]$ have three equally spaced and symmetric support points over the grid of σ_u and σ_v values. The optimum design weights were symmetric over the symmetric design region $\mathscr{X} = [-1, 1]$ and asymmetric over the asymmetric design region $\mathscr{X} = [0, 1]$. This was expected since the trace criterion function is not invariant to translational transformations of the design space.

7.7 Approximations of the Information Matrix

The information matrix for the asymmetrically distributed statistical model (7.1) involves expectations of complicated functions, which makes evaluation of the information matrix difficult. A solution to this problem is to approximate the information matrix. Three approximation methods were presented in Chapter 3. Method 1, the recommended method, approximates the information matrix defined by the first-order derivatives of the log-likelihood function, definition (D.2), by approximating the first-order derivatives. Method 2 also uses definition (D.2) of the information matrix, but in this case, the approximation is carried out by approximating the *product* of the first-order derivatives. Method 3 approximates the information matrix by approximating the second-order derivatives in definition (D.3) of the information matrix. The first method is recommended, and was used in the numerical examples, since it guarantees nonnegative definiteness of the information matrix, whereas the latter two methods do not.

Properties of the information matrix under approximation Method 1 were also given. The approximated information matrix under approximation Method 1, given in equation (6.1), has the same structure as the exact information matrix, given in equation (6.4). The similarity in the structure is such that they have the same rank for any partitioning of the information matrix. Hence any theoretical results based on the rank of the information matrix apply to both

the approximated and exact information matrix. The difference between the two matrices are in the complicated functions (or expectations of functions), which are functions of σ_u and σ_v. Since the optimum design depends on the $\boldsymbol{\tau}$ parameters, which are functions of σ_u and σ_v, the optimum design found using a numerical algorithm, such as Torsney's (1977) algorithm, may be sensitive to an approximation of the information matrix. As with the exact information matrix, optimum designs with approximated information matrices are independent of $\boldsymbol{\beta}$ but depend on $\boldsymbol{\tau}$.

Note that, if approximation Methods 2 or 3 are used, the information matrix is not symmetric. Consequently, a 'correction' to the Gâteaux derivatives, used in the General Equivalence Theorem and in finding the optimising design weights, is required. However, these two approximation methods are not recommended.

7.8 Further Work

7.8.1 Nonlinear models

The focus in this dissertation was primarily on linear and nonlinear optimum design problems for the *linear* model (7.2) with skewed composed error, with numerical examples given for the special case of the log-linear stochastic production frontier model. It was demonstrated that, for these linear models, the optimum design may be dependent on the variance parameters. Only general comments about the parameter dependence of designs for nonlinear model (7.2) were made. Further insight into the affects of an asymmetric composed error on optimum designs can be gained by extending the work presented here to nonlinear models.

7.8.2 Sensitivity to approximation methods

Since the optimum design may depend on the approximation method applied to the information matrix, it would be beneficial to carry out an assessment of the effect of approximation methods. This could be achieved through a simulation study. Rather than approximating the information matrix using Method 1 described in this dissertation, another method of approximation would be to numerically evaluate the complicated functions appearing in the information matrix. This would require a good choice of quadrature method for numerically evaluating integrals which form part of these functions.

7.8.3 Sensitivity to distributional assumptions

(Ritter & Simar 1997) propose that, for estimation of efficiency, the normal-half normal or normal-exponential error specifications should be used because of their simplicity compared to the more flexible normal-truncated normal and normal-gamma specifications. Their argument is based on the assertion that the rankings of the efficiencies are not sensitive to the distributional assumption, even between the normal-half normal and normal-exponential specifications. However, it was shown that the choice of distributional assumption can affect the optimum design, thus potentially affecting the precision of the model estimates. It is possible to carry out a simulation study to investigate the affects of the distributional assumption on the optimum design with respect to the precision of the parameter estimates.

7.8.4 Sensitivity to choices of linear combinations

It is also possible to carry out a simulation study on the affects of the choice of the vector $\boldsymbol{a}$ in the linear combination $\boldsymbol{a}^T\tilde{\boldsymbol{\tau}}$ under a specific distributional assumption. For example, under the normal-half normal error specification, there

were two possible choices for the vector $\boldsymbol{a}$ which produced optimum designs with the same points of support but with different design weights.

7.8.5 A linear combination for precise estimation of efficiency

The formula for calculating the average efficiency over all organisations, or for each individual organisation, is nonlinear. Linearisation of this formula, by use of a Taylor approximation for example, provides a method for designing optimally for estimation of economic efficiency. The linearised formula can be used to determine the matrix A in the linear combination $A^T\boldsymbol{\theta}$. However, it would be difficult to assess the affects of the approximation in the linearisation of the formula, compounded with the approximation of the information matrix.

It might be hoped that an optimum design for good estimation of the $\boldsymbol{\beta}$ parameters would give an optimum estimate of economic efficiency, since calculation of the efficiency requires the $\boldsymbol{\beta}$ parameters. However, the calculation for efficiency also requires the variance parameters and it is not possible to design optimally for both variance parameters since the information matrix is rank deficient. Optimal estimation of the model parameters may be the best that can be achieved.

7.8.6 Other types of frontier models

The theories and methods presented in this dissertation can be extended for other types of frontier models, of which some were briefly discussed in Chapter 4. For example, an extension to the cross-sectional log-linear stochastic *cost* frontier is fairly straight forward. The error for the production frontier is $E = V - U$ and for the cost frontier it is $E = V + U$. Hence this extension requires some changes in signs which carry through all the equations.

In Section 4.4.2 it was noted that stochastic production frontier models are

just random effects or variance component models. Optimum designs have been investigated in the literature for variance component models when the random effects have zero mean. This body of work could be extended by considering a variance component model where one of the random effects has nonzero mean, such as in a longitudinal time-invariant stochastic production frontier model.

Clearly there is scope for future development on optimum designs for models with skewed composed error. The suggestions given above are a selection of some of the possibilities.

Bibliography

Abramowitz, M. & Stegun, I. A. (1965), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York.

Afriat, S. N. (1972), 'Efficiency estimation of production functions', *Int. Econ. Rev.* **13**(3), 568–598.

Aigner, D. J. & Balestra, P. (1988), 'Optimal experimental design for error component models', *Econometrica* **56**(4), 955–971.

Aigner, D. J. & Chu, S. F. (1968), 'On estimating the industry production function', *Amer. Econ. Rev.* **58**(4), 826–839.

Aigner, D., Lovell, C. A. K. & Schmidt, P. (1977), 'Formulation and estimation of stochastic frontier production function models', *J. Econometrics* **6**(1), 21–37.

Alahmadi, A. M. (1993), Algorithms for the Construction of Constrained and Uncontrained Optimal Designs, PhD thesis, University of Glasgow.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Mathematical Statistics, 2nd edn, Wiley, New York.

Atkinson, A. C. (1996), 'The usefulness of optimum experimental designs', *J. R. Statist. Soc. B* **58**(1), 59–76.

Atkinson, A. C. (2008), 'Examples of the use of an equivalence theorem in constructing optimum experimental designs for random-effects nonlinear regression models', *J. Statist. Plann. Inference* (article in press). doi: 10.1016/j.jspi.2008.03.002.

Atkinson, A. C. & Bailey, R. A. (2001), 'One hundred years of the design of experiments on and off the pages of biometrika', *Biometrika* **88**(1), 53–97.

Atkinson, A. C., Bogacka, B. & Zhigljavsky, A. A., eds (2001), *Optimum Design 2000*, Vol. 51 of *Nonconvex Optimization and its Applications*, Kluwer Academic Publishers, the Netherlands.

Atkinson, A. C. & Donev, A. N. (1992), *Optimum Experimental Designs*, Oxford Statistical Science Series 8, Clarendon Press, Oxford. Reprint with corrections.

Atkinson, A. C., Donev, A. N. & Tobias, R. (2007), *Optimum Experimental Designs, with SAS*, Oxford Statistical Science Series 34, Oxford University Press, Oxford.

Atkinson, A. C. & Fedorov, V. V. (1989), *Encyclopedia of Statistical Sciences*, Vol. Supplement, Wiley, New York, chapter Optimum Design of Experiments, pp. 107–114.

Atkinson, T. (2005), Measurement of government output and productivity for the national accounts, Atkinson review: Final report, HM Treasury, http://www.statistics.gov.uk/about/data/methodology/specific/PublicSector/... ...atkinson/downloads/Atkinson_Report_Full.pdf.

Battese, G. E. & Coelli, T. J. (1988), 'Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data', *J. Econometrics* **38**(3), 387–299.

Battese, George, E. & Coelli, T. J. (1992), 'Frontier production functions, technical efficiency and panel data: with aplication to paddy farmers in India', *J. Productiv. Anal.* **3**(1-2), 153–169.

Baum, L. E. & Eagon, J. A. (1967), 'An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology', *Bull. Amer. Math. Soc.* **73**(3), 360–363.

Beckers, D. E. & Hammond, C. J. (1987), 'A tractable likelihood function for the normal-gamma stochastic frontier model', *Econ. Letters* **24**(1), 33–38.

Ben-Israel, A. & Greville, T. N. E. (1974), *Generalized Inverses, Theory and Applications*, Wiley, New York.

Bera, A. K. & Sharma, S. C. (1999), 'Estimating production uncertainty in stochastic frontier production function models', *J. Productiv. Anal.* **12**(3), 187–210.

Bhatia, R. (2007), *Positive Definite Matrices*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton.

Chaloner, K. & Verdinelli, I. (1995), 'Bayesian experimental design: a review', *Statist. Sci.* **10**(3), 273–304.

Chang, F.-C. (1999), 'Exact D-optimal designs for polynomial regression without intercept', *Statist. Probab. Lett.* **44**(2), 131–136.

Charnes, A., Cooper, W. W., Lewin, A. Y. & Seiford, L. M. (1994), *Data Envelopment Analysis: Theory, Methodology and Application*, Kluwer Academic, London.

Charnes, A., Cooper, W. W. & Rhodes, E. (1978), 'Measuring the efficiency of decision making units', *European J. Oper. Res.* **2**(6), 429–444.

Cobb, C. & Douglas, P. H. (1928), 'A theory of production', *Amer. Econ. Rev.* **18**(Supplement), 139–165.

Coelli, T. J. (1995), 'Estimators and hypothesis tests for a stochastic frontier function: A monte carlo analysis', *J. Productiv. Anal.* **6**(3), 247–268.

Coelli, T. J. (n.d.), A guide to frontier version 4.1: A computer program for stochastic frontier production and cost function estimation. Centre for Efficiency and Productivity Analysis (CEPA) Working Paper.

Cooper, W. W., Seiford, L. M. & Tone, K. (2000), *Data Envelopment Analysis: A comprehensive text with models, applications, references, and DEA-Solver software*, Kluwer Academic, London.

Cooper, W. W., Seiford, L. M. & Zhu, J. (2004), *Handbook on Data Envelopment Analysis*, International series in operations research & science; 71, Kluwer Academic, London.

Cornwell, C., Schmidt, P. & Sickles, R. C. (1990), 'Production frontiers with cross-sectional and time-series variations in efficiency levles.', *J. Econometrics* **46**(1-2), 185–200.

Debreu, G. (1951), 'The coefficient of resource utilization', *Econometrica* **19**(3), 273–292.

Di Bucchianico, A., Läuter, H. & Wynn, H. P., eds (2004), *mODa 7 - Advances in Model-Oriented Design and Analysis: Proceedings of the 7th International Workshop on Model-Oriented Design and Analysis held in Heeze, the Netherlands, June 14-18, 2004*, Contributions to Statistics, Physica-Verlag, Heidelberg.

Dodge, Y., Fedorov, V. V. & Wynn, H. P., eds (1988), *Optimal Design and Analysis of Experiments*, North-Holland, the Netherlands.

Elandt-Johnson, R. C. & Johnson, N. L. (1980), *Survival Models and Data Analysis*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Farrell, M. J. (1957), 'The measurement of productive efficiency', *J. R. Stat. Soc. Ser. A. Gen.* **120**(3), 253–281.

Fedorov, V. V. (1972), *Theory of Optimal Experiments*, Probability and Mathematical Statistics 12, Academic Press, New York; London.

Fedorov, V. V. (1978), 'Properties of optimal designs of regression experiments in singular cases', *Translated from Staticheskie Metody* pp. 151–153.

Fedorov, V. V. & Läuter, H., eds (1987), *Model-Oriented Data Analysis: proceedings of an IIASA (International Institute for Applied Systems Analysis) workshop on data analysis held at Eisenach, GDR, March 9-13, 1987*, Vol. 297 of *Lecture Notes in Economics and Mathematical Systems*, Springer-Verlag.

Fellman, J. (1974), 'On the allocation of linear observations', *Soc. Sci. Fenn. Coment. Phys-Math* **44**(1), 27–28.

Ford, I. & Silvey, S. D. (1980), 'A sequentially constructed design for estimating a nonlinear parametric function', *Biometrika* **67**(2), 381–388.

Fried, H. O., Lovell, C. A. K. & Schmidt, S. S., eds (1993), *The Measurement of Productive Efficiency: Techniques and Applications*, Oxford University Press, New York.

Gershon, P. (2004), Releasing resources to the frontline, Independent review of public sector efficiency, HM Treasury, http://www.hm-treasury.gov.uk./media/C/A/efficiency_review120704.pdf.

Giovagnoli, A. & Sebastiani, P. (1989), 'Experimental designs for mean and variance estimation in variance component models', *Comput. Statist. Data Anal.* **8**(1), 21–28.

Greene, W. H. (1980*a*), 'Maximum likelihood estimation of econometric frontier functions', *J. Econometrics* **13**(1), 27–56.

Greene, W. H. (1980*b*), 'On the estimation of a flexible frontier production model', *J. Econometrics* **13**(1), 101–115.

Greene, W. H. (1990), 'A gamma-distributed stochastic frontier model', *J. Econometrics* **46**(1–2), 141–163.

Happacher, M. (1995), 'Exact and approximate D-optimal designs in polynomial regression', *Metrika* **42**(1), 19–27.

Healy, M. (2000), *Matrices for Statistics*, 2nd edn, Clarendon Press, Oxford.

Hjalmarsson, L., Kumbhakar, S. C. & Heshmati, A. (1996), 'DEA, DFA and SFA: A comparison', *J. Productiv. Anal.* **7**(2-3), 303–327.

HM Treasury (2004), Stability, security and opportunity for all: Investing for Britain's long-term future, 2004 Spending Review: New Public Speding Plans 2005-2008 CM6237, HM Treasury, http://www.hm-treasury.gov.uk/spending_review/spend_sr04/report/spend_sr04_repindex.cfm.

Horrace, W. C. & Schmidt, P. (1996), 'Confidence statements for efficiency estimates from stochastic frontier models', *J. Productiv. Anal.* **7**(2-3), 257–282.

Jacobs, R., Smith, P. C. & Street, A. (2006), *Measuring Efficiency in Health Care: Analytic Techniques and Health Policy*, Cambridge University Press, Cambridge.

Johnson, N. L. & Kotz, S. (1970), *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 1, Houghton Mifflin Company, New York.

Jondrow, J., Lovell, C. A. K., Materov, I. S. & Schmidt, P. (1982), 'On the estimation of technical inefficiency in the stochastic frontier production function model', *J. Econometrics* **19**(2-3), 233–238.

Kalirajan, K. P. & Shand, R. T. (1999), 'Frontier production functions and technical efficiency measures', *J. Econ. Surveys* **13**(2), 149–172.

Khuri, A. I. (2000), 'Dsigns for variance components estimation: past and present', *Int. Statistical Rev.* **68**(3), 311–322.

Kiefer, J. (1974), 'General equivalence theorem for optimum designs (approximate theory)', *Ann. Statist.* **2**(5), 849–879.

Koopmans, T. C. (1951), An analysis of production as efficient combination of activities, *in* T. C. Koopmans, ed., 'Activity Analysis of Production and Allocation', Cowles Commission for Research in Economics, Monograph 13, Wiley, New York.

Kumbhakar, S. C. (1987), 'The specification of technical and allocative inefficiency of multi-product firms in stochastic production and profit frontiers', *Journal of Quantitative Economics (a biannual Journal of the Indian Econometric Society)* **3**, 213–223.

Kumbhakar, S. C. (1990), 'Production frontiers, panel data and time-varying technical inefficiency', *J. Econometrics* **46**(1-2), 201–211.

Kumbhakar, S. C. & Lovell, C. A. K. (2000), *Stochastic Frontier Analysis*, Cambridge University Press, Cambridge.

Lee, L.-F. & Tyler, W. G. (1978), 'The stochastic frontier production function and average efficiency : An empirical analysis', *J. Econometrics* **7**(3), 385–389.

Lee, Y. H. & Schmidt, P. (1993), *A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency*, Oxford University Press, New York, chapter 8, pp. 237–255.

Lohr, S. L. (1995), 'Optimal bayesian design of experiments for the one-way random effects model', *Biometrika* **82**(1), 175–186.

Mandal, S. & Torsney, B. (2000), 'Algorithms for the construction of optimizing distributions', *Comm. Statist. Theory Methods* **29**, 1219–1231.

Mandal, S. & Torsney, B. (2006), 'Construction of optimal designs using a clustering approach', *J. Statist. Plann. Inference* **136**(3), 1120–1134.

Meeusen, W. & van den Broeck, J. (1977), 'Efficiency estimation from Cobb-Douglas production functions with composed error', *Int. Econ. Rev.* **18**(2), 435–444.

Melas, V. B. (2006), *Functional Approach to Optimal Experimental Design*, number 184 *in* 'Lecture Notes in Statistics', Springer, New York.

Mentre, F., Mallet, A. & Baccar, D. (1997), 'Optimal design in random-effects regression models', *Biometrika* **84**(2), 429–442.

Moore, E. H. (1920), 'On the reciprocal of the general algebraic matrix', *Bull. Amer. Math. Soc.* **26**, 394–395.

Morrell, C. H. (1998), 'Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood', *Biometrics* **54**(4), 1560–1568.

Mukerjee, R. & Huda, S. (1988), 'Optimal design for the estimation of variance components', *Biometrika* **75**(1), 75–80.

Müller, W. G. (2001), *Collecting Spatial Data: Optimum Design of Experiments for Random Fields*, Contributions to Statistics, 2nd revised edn, Physica-Verlag, Heidelberg.

Murillo-Zamorano, L. R. (2004), 'Economic efficiency and frontier techniques', *J. Econ. Surveys* **18**(1), 33–77.

Nakamura, T. (1980), 'On the moment of positively truncated normal distribution', *Journ. Japan Statist. Soc.* **10**(2), 139–144.

Pázman, A. (1978), 'Computation of the optimum designs under singular information matrices', *Ann. Statist.* **6**(2), 465–467.

Penrose, R. (1955), 'A generalized inverse for matrices', *Math. Proc. Cambridge Philos. Soc.* **51**, 406–413.

Pitt, M. M. & Lee, L.-F. (1981), 'The measurement and sources of technical inefficiency in the Indonesian weaving industry', *J. Devel. Econ.* **9**(1), 43–64.

Pukelsheim, F. (1980), 'On linear regression designs wihich maximize information', *J. Statist. Plann. Inference* **4**(4), 339–364.

Pukelsheim, F. (1993), *Optimal Design of Experiments*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Pukelsheim, F. & Torsney, B. (1991), 'Optimal weights for experimental designs on linearly independent support points', *Ann. Statist.* **19**(3), 1614–1625.

Richmond, J. (1974), 'Estimating the efficiency of production', *Int. Econ. Rev.* **15**(2), 515–521.

Ritter, C. & Simar, L. (1997), 'Pitfalls of normal-gamma stochastic frontier models', *J. Productiv. Anal.* **8**(2), 167–182.

Rohde, C. A. (1965), 'Generalized inverses of partitioned matrices', *SIAM J. Appl. Math.* **13**(4), 1033–1035.

Rohde, C. A. (1966), 'Some results on generalized inverses', *SIAM Review* **8**(2), 201–205.

Schervish, M. J. (1995), *Theory of Statistics*, Springer, New York.

Schmidt, P. (1976), 'On the statistical estimation of parametric frontier production functions', *Rev. Econ. Statist.* **58**(2), 238–239.

Schmidt, P. & Lin, T.-F. (1984), 'Simple tests of alternative specifications in stochastic frontier models', *J. Econometrics* **24**(3), 349–361.

Schmidt, P. & Sickles, R. C. (1984), 'Production frontiers and panel data', *J. Bus. Econ. Statist.* **2**(4), 367–374.

Sena, V. (1999), 'Stochastic frontier estimation: A review of the software options', *J. Appl. Econometrics* **14**(5), 579–586.

Silvey, S. D. (1978), 'Optimal design measures with singular information matrices', *Biometrika* **65**(3), 553–559.

Silvey, S. D. (1980), *Optimal Design : An Introduction to the Theory for Parameter Estimation*, Monographs on Applied Probability and Statistics, Chapman and Hall, London.

Silvey, S. D., Titterington, D. M. & Torsney, B. (1978), 'An algorithm for optimal designs on a finite design space', *Comm. Statist. Theory Methods* **14**, 1379–1389.

Stevenson, R. E. (1980), 'Likelihood functions for generalized stochastic frontier estimation', *J. Econometrics* **13**(1), 57–66.

Stewart, J. (1995), *Calculus*, 3rd edn, Brooks/Cole, USA.

Stram, D. O. & Lee, J. W. (1994), 'Variance components testing in the longitudinal mixed effects model', *Biometrics* **50**(4), 1171–1177.

Torsney, B. (1977), 'Contribution to discussion on the paper "Maximum likelihood from incomplete data via the EM algorithm" by Dempster et al.', *J. R. Statist. Soc. B* **39**(1), 26–27.

Torsney, B. (1988), Computing optimizing distributions with applications in design, estimation and image processing, *in* Y. Dodge, V. V. Fedorov & H. P. Wynn, eds, 'Optimal Design and Analysis of Experiments : Proceedings of the First International Conference-Workshop on Optimal Design and Analysis of Experiments held in University of Neuchâtel, Switzerland, July 25-28, 1988', Elsevier Science Publishers B.V., Amsterdam: North-Holland, pp. 361–370.

Torsney, B. & Alahmadi, A. M. (1992), Further development of algorithms for constructing optimizing distributions, *in* V. V. Fedorov, W. G. Müller & I. N. Vuchkov, eds, 'Model Oriented Data-Analysis: A Survey of Recent Methods : Proceedings of the 2nd IIASA-Workshop in St. Kyrik, Bulgaria, May 28-June 1, 1990', Contributions to Statistics, Physica-Verlag, Heidelberg, pp. 121–129.

Torsney, B. & Mandal, S. (2001), Construction of constrained optimal designs, *in* A. C. Atkinson, B. Bogacka & A. A. Zhigljavsky, eds, 'Optimum Design 2000: Papers presented at the conference "Optimum Design 2000: Prospects for the New Millennium" held in Cardiff, United Kingdom, April 12-14, 2000', Vol. 51 of *Nonconvex Optimization and its Applications*, Kluwer Academic Publishers, the Netherlands, pp. 141–152.

Torsney, B. & Mandal, S. (2004), Multiplicative algorithms for constructing optimizing distributions: further developments, *in* A. Di Bucchianico, H. Läuter & H. P. Wynn, eds, 'mODa 7 - Advances in Model-Oriented Design and Analysis: Proceedings of the 7th International Workshop on Model-Oriented Design and Analysis held in Heeze, the Netherlands, June 14-18, 2004', Contributions to Statistics, Physica-Verlag, Heidelberg, pp. 163–171.

Van den Broeck, J., Koop, G., Osiewalski, J. & Steel, M. F. J. (1994), 'Stochastic frontier models: A bayesian perspective', *J. Econometrics* **61**(2), 273–303.

Watt, S. M. (2006), Pivot-free block matrix inversion, *in* W. Decker, M. Dewar, E. Kaltofen & S. Watt, eds, 'Challenges in Symbolic Computation Software', number 06271 *in* 'Dagstuhl Seminar Proceedings', Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, Dagstuhl, Germany. http://drops.dagstuhl.de/opus/volltexte/2006/780.

White, L. V. (1973), 'An extension of the general equivalence theorem to nonlinear models', *Biometrika* **60**(2), 345–348.

Whittle, P. (1973), 'Some general points in the theory of optimal experimental design', *J. R. Statist. Soc. B* **35**(1), 123–130.

Winsten, C. B. (1957), 'Discussion on Mr. Farrell's paper', *J. R. Stat. Soc. Ser. A. Gen.* **120**(3), 282–284.

Wynn, H. P. (1972), 'Results in the theory and construction of D-optimum experimental designs', *J. R. Statist. Soc. B* **34**(2), 133–147.

Wynn, H. P. (1984), 'Jack Kiefer's contributions to experimental design', *Ann. Statist.* **12**(2), 416–423.

Appendix A

Derivation of Information Matrices for the General Model

A.1 Calculations for the Normal-Exponential Model

The detailed calculations for deriving the information matrix for the normal-exponential model in Section 2.3 of Chapter 2 are given here.

The probability density function of $U \sim \text{Exponential}(1/\sigma_u)$ is

$$f_U(u;\sigma_u) = \frac{1}{\sigma_u}\exp\left\{-\frac{u}{\sigma_u}\right\}, \qquad u \geq 0,\ \sigma_u > 0, \tag{A.1}$$

with mean and variance

$$\begin{aligned} \mathbb{E}[U] &= \sigma_u, \\ Var(U) &= \sigma_u^2. \end{aligned}$$

The density of V is given in equation (2.4) with mean and variance given in equation (2.5). The joint probability density function of U and V is

$$f_{U,V}(u,v) = f_U(u) \cdot f_V(v)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_u\sigma_v}\exp\left\{-\frac{u}{\sigma_u}-\frac{v^2}{2\sigma_v^2}\right\}. \tag{A.2}$$

The joint density function of U and $E = c_uU + c_vV$ is given by

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= \frac{1}{|c_v|}f_{U,V}\left(u,\frac{\varepsilon-c_uu}{c_v}\right) \\ &= \frac{1}{|c_v|\sqrt{2\pi}\sigma_u\sigma_v}\exp\left\{-\frac{u}{\sigma_u}-\frac{(\varepsilon-c_uu)^2}{2c_v^2\sigma_v^2}\right\} \\ &= \frac{1}{|c_v|\sqrt{2\pi}\sigma_u\sigma_v}\exp\left\{-\frac{1}{2}\left[\frac{c_u^2}{c_v^2\sigma_v^2}u^2-2\left(\frac{c_u\varepsilon}{c_v^2\sigma_v^2}-\frac{1}{\sigma_u}\right)u+\frac{\varepsilon^2}{c_v^2\sigma_v^2}\right]\right\}. \end{aligned} \tag{A.3}$$

If we let $K = \frac{1}{|c_v|\sqrt{2\pi}\sigma_u\sigma_v}$, $A = \frac{c_u^2}{c_v^2\sigma_v^2}$, $B = \frac{c_u\varepsilon}{c_v^2\sigma_v^2}-\frac{1}{\sigma_u}$ and $C = \frac{\varepsilon^2}{c_v^2\sigma_v^2}$ then the joint density function of U and E becomes

$$f_{U,E}(u,\varepsilon) = K\exp\left\{-\frac{1}{2}\left[Au^2-2Bu+C\right]\right\}.$$

When the joint density of U and E is of this form, the marginal density of E is given by equation (C.6) as

$$f_E(\varepsilon) = K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B}{\sqrt{A}}\right),$$

where

$$K\sqrt{\frac{2\pi}{A}} = \frac{1}{|c_u|\sigma_u},$$

$$C-\frac{B^2}{A} = \frac{2\varepsilon}{c_u\sigma_u}-\frac{c_v^2\sigma_v^2}{c_u^2\sigma_u^2},$$

$$\frac{B}{\sqrt{A}} = \frac{c_u\varepsilon}{|c_uc_v|\sigma_v}-\frac{|c_v|\sigma_v}{|c_u|\sigma_u}.$$

The marginal density of E is then given by

$$f_E(\varepsilon) = \frac{1}{|c_u|\sigma_u}\exp\left\{-\frac{\varepsilon}{c_u\sigma_u}+\frac{c_v^2\sigma_v^2}{2c_u^2\sigma_u^2}\right\}\Phi\left(\frac{c_u\varepsilon}{|c_uc_v|\sigma_v}-\frac{|c_v|\sigma_v}{|c_u|\sigma_u}\right),$$

with mean and variance that can be derived using equations (C.7) and (C.8) in Appendix C and which are given by

$$\begin{aligned} \mathbb{E}[E] &= c_u\sigma_u, \\ Var(E) &= c_u^2\sigma_u^2 + c_v^2\sigma_v^2. \end{aligned}$$

A.1.1 Log-likelihood function

The log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_v)$ is given in equation (2.17). Reparameterising the term $\varepsilon = y - f(\boldsymbol{x}, \boldsymbol{\beta})$ as a function of the variable a, the first-order derivatives of $\ln f_Y(y; \boldsymbol{\theta})$ are

$$\begin{aligned} \frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= -\left\{-\frac{1}{c_u\sigma_u} + \frac{c_u}{|c_uc_v|\sigma_v}h(a)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \\ \frac{\partial \ln f_Y}{\partial(1/\sigma_u)} &= \sigma_u - \frac{y - f(\boldsymbol{x},\boldsymbol{\beta})}{c_u} + \frac{c_v^2\sigma_v^2}{c_u^2\sigma_u} - \frac{|c_v|\sigma_v}{|c_u|}h(a) \\ &= \sigma_u + \frac{|c_v|\sigma_v}{|c_u|}[a - h(a)], \\ \frac{\partial \ln f_Y}{\partial \sigma_v^2} &= \frac{c_v^2}{2c_u^2\sigma_u^2} - \left(\frac{c_u[y - f(\boldsymbol{x},\boldsymbol{\beta})]}{2|c_uc_v|\sigma_v^3} + \frac{|c_v|}{2|c_u|\sigma_u\sigma_v}\right)h(a) \\ &= \frac{c_v^2}{2c_u^2\sigma_u^2} - \left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v} - \frac{1}{2\sigma_v^2}a\right)h(a). \end{aligned}$$

The corresponding second-order derivatives are

$$\begin{aligned} \frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \left\{\frac{1}{c_v^2\sigma_v^2}h(a)[a - h(a)]\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T \\ &\quad - \left\{-\frac{1}{c_u\sigma_u} + \frac{c_u}{|c_uc_v|\sigma_v}h(a)\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \\ &= \left\{\frac{1}{c_v^2\sigma_v^2}[ah(a) - h(a)^2]\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T \end{aligned}$$

$$-\left\{-\frac{1}{c_u\sigma_u}+\frac{c_u}{|c_uc_v|\sigma_v}h(a)\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)^2} &= -\sigma_u^2+\left(\frac{|c_v|\sigma_v}{|c_u|}\right)^2+\left(\frac{|c_v|\sigma_v}{|c_u|}\right)^2 h(a)[a-h(a)]\\ &= -\sigma_u^2+\left(\frac{|c_v|\sigma_v}{|c_u|}\right)^2[1+ah(a)-h(a)^2],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial(\sigma_v^2)^2} &= \left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v^3}-\frac{3}{4\sigma_v^4}a\right)h(a)+\left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v}-\frac{1}{2\sigma_v^2}a\right)^2 h(a)[a-h(a)]\\ &= \left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v}\right)^2[ah(a)-h(a)^2]-\frac{|c_v|}{|c_u|\sigma_u\sigma_v^3}[a^2h(a)-ah(a)^2-h(a)]\\ &\quad+\frac{1}{4\sigma_v^4}[a^3h(a)-a^2h(a)^2-3ah(a)],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)} &= \left\{\frac{1}{c_u}+\frac{1}{c_u}h(a)[a-h(a)]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\\ &= \frac{1}{c_u}\left\{1+ah(a)-h(a)^2\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_v^2} &= \left\{\frac{c_u}{2|c_uc_v|\sigma_v^3}h(a)+\left(\frac{1}{c_u\sigma_u\sigma_v^2}-\frac{c_u}{2|c_uc_v|\sigma_v^3}a\right)h(a)[a-h(a)]\right\}\times\\ &\quad\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\\ &= \left\{\frac{c_u}{2|c_uc_v|\sigma_v^3}[h(a)-a^2h(a)+ah(a)^2]\right.\\ &\quad\left.+\frac{1}{c_u\sigma_u\sigma_v^2}[ah(a)-h(a)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)\partial\sigma_v^2} &= \frac{c_v^2}{c_u^2\sigma_u}-\frac{|c_v|}{2|c_u|\sigma_v}h(a)+\left(\frac{c_v^2}{c_u^2\sigma_u}-\frac{|c_v|}{2|c_u|\sigma_v}a\right)h(a)[a-h(a)]\\ &= \frac{c_v^2}{c_u^2\sigma_u}[1+ah(a)-h(a)^2]-\frac{|c_v|}{2|c_u|\sigma_v}[h(a)+a^2h(a)-ah(a)^2].\end{aligned}$$

A.1.2 Information matrix in terms of first-order partial derivatives

The form of the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_v)$ is given in equation (2.18). Dispensing with the observation subscripts, the components of the information matrix are

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)^T\right] = \left\{\left(\frac{1}{c_u\sigma_u}\right)^2 - \frac{2}{|c_u c_v|\sigma_u\sigma_v}\mathbb{E}[h(a)] + \left(\frac{1}{|c_v|\sigma_v}\right)^2 \mathbb{E}[h(a)^2]\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T,$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)^2\right] = \sigma_u^2 + 2\frac{|c_v|\sigma_u\sigma_v}{|c_u|}\left(\mathbb{E}[a] - \mathbb{E}[h(a)]\right) + \left(\frac{|c_v|\sigma_v}{|c_u|}\right)^2\left(\mathbb{E}[a^2] - 2\mathbb{E}[ah(a)] + \mathbb{E}[h(a)^2]\right),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)^2\right] = \left(\frac{c_v^2}{2c_u^2\sigma_u^2}\right)^2 - \frac{|c_v|^3}{|c_u|^3\sigma_u^3\sigma_v}\mathbb{E}[h(a)] + \frac{c_v^2}{2c_u^2\sigma_u^2\sigma_v^2}\mathbb{E}[ah(a)] + \left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v}\right)^2\mathbb{E}[h(a)^2] - \frac{|c_v|}{|c_u|\sigma_u\sigma_v^3}\mathbb{E}[ah(a)^2] + \left(\frac{1}{2\sigma_v^2}\right)^2\mathbb{E}[a^2h(a)^2],$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)\right] = \left\{\frac{1}{c_u}\left(1 - \mathbb{E}[ah(a)] + \mathbb{E}[h(a)^2]\right) + \frac{|c_v|\sigma_v}{c_u^2\sigma_u}\left(\mathbb{E}[a] - \mathbb{E}[h(a)]\right) - \frac{c_u\sigma_u}{|c_u c_v|\sigma_v}\mathbb{E}[h(a)]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] &= \left\{\frac{c_v^2}{2c_u^3\sigma_u^3} - \frac{|c_v|}{|c_u|c_u\sigma_u^2\sigma_v}\mathbb{E}[h(a)]\right. \\ &+\frac{1}{2c_u\sigma_u\sigma_v^2}\mathbb{E}[ah(a)] - \frac{|c_v|}{2|c_u|c_u\sigma_u^2\sigma_v}\mathbb{E}[h(a)] + \frac{1}{c_u\sigma_u\sigma_v^2}\mathbb{E}[h(a)^2] \\ &\left.-\frac{c_u}{2|c_uc_v|\sigma_v^3}\mathbb{E}[ah(a)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] &= \frac{c_v^2}{2c_u^2\sigma_u} - \frac{|c_v|}{|c_u|\sigma_v}\mathbb{E}[h(a)] + \frac{\sigma_u}{2\sigma_v^2}\mathbb{E}[ah(a)] \\ &+\frac{|c_v|^3\sigma_v}{2|c_u|^3\sigma_u^2}\left(\mathbb{E}[a] - \mathbb{E}[h(a)]\right) - \frac{c_v^2}{c_u^2\sigma_u}\mathbb{E}[ah(a)] + \frac{|c_v|}{2|c_u|\sigma_v}\mathbb{E}[a^2h(a)] \\ &+\frac{c_v^2}{c_u^2\sigma_u}\mathbb{E}[h(a)^2] - \frac{|c_v|}{2|c_u|\sigma_v}\mathbb{E}[ah(a)^2].\end{aligned}$$

A.1.3 Information matrix in terms of second-order partial derivatives

An alternative formulation for the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_u, \sigma_v)$ is given in equation (2.19). The components of the per observation expected Fisher information matrix are

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right] &= -\left\{\frac{1}{c_v^2\sigma_v^2}\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right)\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right) \\ &+\left\{-\frac{1}{c_u\sigma_u} + \frac{c_u}{|c_uc_v|\sigma_v}\mathbb{E}[h(a)]\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T},\end{aligned}$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial (1/\sigma_u)^2}\right] = \sigma_u^2 - \left(\frac{|c_v|\sigma_v}{|c_u|}\right)^2\left(1 + \mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right),$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial (\sigma_v^2)^2}\right] &= -\left(\frac{|c_v|}{|c_u|\sigma_u\sigma_v}\right)^2\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right) \\ &+\frac{|c_v|}{|c_u|\sigma_u\sigma_v^3}\left(\mathbb{E}[a^2h(a)] - \mathbb{E}[ah(a)^2] - \mathbb{E}[h(a)]\right)\end{aligned}$$

$$-\frac{1}{4\sigma_v^4}\left(\mathbb{E}[a^3h(a)] - \mathbb{E}[a^2h(a)^2] - \mathbb{E}[3ah(a)]\right),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)}\right] = -\frac{1}{c_u}\left\{1 + \mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_v^2}\right] &= -\left\{\frac{c_u}{2|c_uc_v|\sigma_v^3}\left(\mathbb{E}[h(a)] - \mathbb{E}[a^2h(a)] + \mathbb{E}[ah(a)^2]\right)\right.\\ &\quad\left.+\frac{1}{c_u\sigma_u\sigma_v^2}\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)\partial\sigma_v^2}\right] &= -\frac{c_v^2}{c_u^2\sigma_u}\left(1 + \mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right)\\ &\quad+\frac{|c_v|}{2|c_u|\sigma_v}\left(\mathbb{E}[h(a)] + \mathbb{E}[a^2h(a)] - \mathbb{E}[ah(a)^2]\right).\end{aligned}$$

A.2 Calculations for the Normal-Truncated Normal Model

The detailed calculations for deriving the information matrix for the normal-truncated normal model in Section 2.4 of Chapter 2 are given below.

The probability density function of $U \sim N^+(\mu, \sigma_u^2)$ is

$$f_U(u;\mu,\sigma_u) = \frac{1}{\sqrt{2\pi}\sigma_u}\exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2}\right\}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1},$$

$$u \geq 0,\ -\infty < \mu < \infty,\ \sigma_u > 0, \tag{A.4}$$

with mean and variance

$$\begin{aligned}\mathbb{E}[U] &= \mu + \phi\left(\frac{\mu}{\sigma_u}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}\sigma_u\\ &= \mu + h\left(-\frac{\mu}{\sigma_u}\right)\sigma_u,\end{aligned}$$

$$\begin{aligned} Var(U) &= \left\{1 - \frac{\mu}{\sigma_u}\phi\left(\frac{\mu}{\sigma_u}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1} - \left[\phi\left(\frac{\mu}{\sigma_u}\right)\right]^2\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-2}\right\}\sigma_u^2 \\ &= \left\{1 - \frac{\mu}{\sigma_u}h\left(-\frac{\mu}{\sigma_u}\right) - \left[h\left(-\frac{\mu}{\sigma_u}\right)\right]^2\right\}\sigma_u^2. \end{aligned}$$

The density of V is given in equation (2.4) with mean and variance given in equation (2.5). The joint probability density function of U and V is

$$\begin{aligned} f_{U,V}(u,v) &= f_U(u)\cdot f_V(v) \\ &= \frac{1}{2\pi\sigma_u\sigma_v}\exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2}\right\}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}. \end{aligned} \quad (A.5)$$

The joint density function of U and $E = c_uU + c_vV$ is given by

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= \frac{1}{|c_v|}f_{U,V}\left(u, \frac{\varepsilon - c_uu}{c_v}\right) \\ &= \frac{1}{|c_v|2\pi\sigma_u\sigma_v}\exp\left\{-\frac{(u-\mu)^2}{2\sigma_u^2} - \frac{(\varepsilon - c_uu)^2}{2c_v^2\sigma_v^2}\right\}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1} \\ &= \frac{1}{|c_v|2\pi\sigma_u\sigma_v}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}\times \\ &\quad \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{\sigma_u^2} + \frac{c_u^2}{c_v^2\sigma_v^2}\right)u^2 - 2\left(\frac{\mu}{\sigma_u^2} + \frac{c_u\varepsilon}{c_v^2\sigma_v^2}\right)u + \frac{\mu^2}{\sigma_u^2} + \frac{\varepsilon^2}{c_v^2\sigma_v^2}\right]\right\}. \end{aligned} \quad (A.6)$$

If we let $K = \frac{1}{|c_v|2\pi\sigma_u\sigma_v}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}$, $A = \frac{1}{\sigma_u^2} + \frac{c_u^2}{c_v^2\sigma_v^2}$, $B = \frac{\mu}{\sigma_u^2} + \frac{c_u\varepsilon}{c_v^2\sigma_v^2}$ and $C = \frac{\mu^2}{\sigma_u^2} + \frac{\varepsilon^2}{c_v^2\sigma_v^2}$ then the joint density function of U and E becomes

$$f_{U,E}(u,\varepsilon) = K\exp\left\{-\frac{1}{2}[Au^2 - 2Bu + C]\right\}.$$

When the joint density of U and E is of this form, the marginal density of E is given by equation (C.6) as

$$f_E(\varepsilon) = K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\}\Phi\left(\frac{B}{\sqrt{A}}\right).$$

If we let

$$\begin{aligned}\sigma_G^2 &= c_u^2\sigma_u^2 + c_v^2\sigma_v^2,\\ \lambda &= \frac{\sigma_u}{\sigma_v},\end{aligned}$$

then

$$\begin{aligned}A &= \frac{\sigma_G^2}{c_v^2\sigma_u^2\sigma_v^2},\\ K\sqrt{\frac{2\pi}{A}} &= \frac{1}{\sqrt{2\pi}\sigma_G}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1},\\ C - \frac{B^2}{A} &= \left(\frac{c_u\mu - \varepsilon}{\sigma_G}\right)^2,\\ \frac{B}{\sqrt{A}} &= \frac{|c_v|\mu}{\lambda\sigma_G} + \frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}.\end{aligned}$$

The marginal density of E is then given by

$$\begin{aligned}f_E(\varepsilon) &= \frac{1}{\sqrt{2\pi}\sigma_G}\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1}\exp\left\{-\frac{1}{2}\left(\frac{c_u\mu-\varepsilon}{\sigma_G}\right)^2\right\}\Phi\left(\frac{|c_v|\mu}{\lambda\sigma_G}+\frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}\right)\\ &= \frac{1}{\sigma_G}\phi\left(\frac{c_u\mu-\varepsilon}{\sigma_G}\right)\Phi\left(\frac{|c_v|\mu}{\lambda\sigma_G}+\frac{c_u\lambda\varepsilon}{|c_v|\sigma_G}\right)\left[\Phi\left(\frac{\mu}{\sigma_u}\right)\right]^{-1},\end{aligned}$$

with mean and variance that can be derived using equations (C.7) and (C.8) in Appendix C and which are given by

$$\begin{aligned}\mathbb{E}[E] &= \tilde{c_u}\sigma_u,\\ Var(E) &= \tilde{\tilde{c_u}}^2\sigma_u^2 + c_v^2\sigma_v^2,\end{aligned}$$

where $\tilde{c_u} = \frac{c_u\mu}{\sigma_u} + c_u h\left(-\frac{\mu}{\sigma_u}\right)$ and $\tilde{\tilde{c_u}}^2 = c_u^2\left\{1 - \frac{\mu}{\sigma_u}h\left(-\frac{\mu}{\sigma_u}\right) - \left[h\left(-\frac{\mu}{\sigma_u}\right)\right]^2\right\}$.

A.2.1 Log-likelihood function

The log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \lambda, \sigma_G)$ is given in equation (2.25). Reparameterising the log-likelihood as a function of the variable a_1, the first-order derivatives of $\ln f_Y(y; \boldsymbol{\theta})$ are

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= -\left\{ \frac{c_u\mu - [y - f(\boldsymbol{x}, \boldsymbol{\beta})]}{\sigma_G^2} + \frac{c_u\lambda}{|c_v|\sigma_G} h(a_1) \right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -\left\{ \frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G^2} + \frac{|c_v|}{c_u\lambda\sigma_G} a_1 + \frac{c_u\lambda}{|c_v|\sigma_G} h(a_1) \right\} \frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \mu} &= -\frac{c_u^2\mu - c_u[y - f(\boldsymbol{x}, \boldsymbol{\beta})]}{\sigma_G^2} + \frac{|c_v|}{\lambda\sigma_G} h(a_1) + \frac{a_2}{\mu} h(a_2) \\
&= -\frac{\mu(c_u^2\lambda^2 + c_v^2)}{\lambda^2\sigma_G^2} - \frac{|c_v|}{\lambda\sigma_G} a_1 + \frac{|c_v|}{\lambda\sigma_G} h(a_1) + \frac{a_2}{\mu} h(a_2),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \lambda} &= \left(-\frac{|c_v|\mu}{\lambda^2\sigma_G} + \frac{c_u[y - f(\boldsymbol{x}, \boldsymbol{\beta})]}{|c_v|\sigma_G} \right) h(a_1) + \frac{c_v^2\mu}{(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G} h(a_2) \\
&= \left(-\frac{2|c_v|\mu}{\lambda^2\sigma_G} - \frac{1}{\lambda} a_1 \right) h(a_1) + \frac{c_v^2\mu}{(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G} h(a_2),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \sigma_G^2} &= -\frac{1}{2\sigma_G^2} + \frac{1}{2}\left(\frac{c_u\mu - [y - f(\boldsymbol{x}, \boldsymbol{\beta})]}{\sigma_G^2} \right)^2 + \frac{1}{2\sigma_G^2} a_1 h(a_1) - \frac{1}{2\sigma_G^2} a_2 h(a_2) \\
&= -\frac{1}{2\sigma_G^2} + \frac{1}{2}\left(\frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G^2} + \frac{|c_v|}{c_u\lambda\sigma_G} a_1 \right)^2 + \frac{1}{2\sigma_G^2} a_1 h(a_1) - \frac{1}{2\sigma_G^2} a_2 h(a_2).
\end{aligned}$$

The corresponding second-order derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \left\{ -\frac{1}{\sigma_G^2} + \left(\frac{c_u\lambda}{|c_v|\sigma_G} \right)^2 h(a_1)[a_1 - h(a_1)] \right\} \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \\
&\qquad - \left\{ \frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G^2} + \frac{|c_v|}{c_u\lambda\sigma_G} a_1 + \frac{c_u\lambda}{|c_v|\sigma_G} h(a_1) \right\} \frac{\partial^2 f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \\
&= \left\{ -\frac{1}{\sigma_G^2} + \left(\frac{c_u\lambda}{|c_v|\sigma_G} \right)^2 [a_1 h(a_1) - h(a_1)^2] \right\} \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \left(\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T
\end{aligned}$$

$$- \left\{ \frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G^2} + \frac{|c_v|}{c_u\lambda\sigma_G} a_1 + \frac{c_u\lambda}{|c_v|\sigma_G} h(a_1) \right\} \frac{\partial^2 f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial\mu^2} &= -\left(\frac{c_u}{\sigma_G}\right)^2 + \left(\frac{|c_v|}{\lambda\sigma_G}\right)^2 h(a_1)[a_1 - h(a_1)] - \left(\frac{a_2}{\mu}\right)^2 h(a_2)[a_2 - h(a_2)] \\
&= -\left(\frac{c_u}{\sigma_G}\right)^2 + \left(\frac{|c_v|}{\lambda\sigma_G}\right)^2 [a_1 h(a_1) - h(a_1)^2] - \left(\frac{a_2}{\mu}\right)^2 h(a_2)[a_2 - h(a_2)],
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial\lambda^2} &= \frac{2|c_v|\mu}{\lambda^3\sigma_G} h(a_1) + \left(-\frac{2|c_v|\mu}{\lambda^2\sigma_G} - \frac{1}{\lambda} a_1\right)^2 h(a_1)[a_1 - h(a_1)] \\
&\quad - \frac{c_v^2\mu(3c_u^2\lambda^2 + 2c_v^2)}{(c_u^2\lambda^2 + c_v^2)^{3/2}\lambda^3\sigma_G} h(a_2) \\
&\quad - \left(\frac{c_v^2\mu}{(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G}\right)^2 h(a_2)[a_2 - h(a_2)] \\
&= \frac{2|c_v|\mu}{\lambda^3\sigma_G} h(a_1) + \left(\frac{2|c_v|\mu}{\lambda^2\sigma_G}\right)^2 [a_1 h(a_1) - h(a_1)^2] \\
&\quad + \frac{4|c_v|\mu}{\lambda^3\sigma_G} [a_1^2 h(a_1) - a_1 h(a_1)^2] + \frac{1}{\lambda^2} [a_1^3 h(a_1) - a_1^2 h(a_1)^2] \\
&\quad - \frac{c_v^2\mu(3c_u^2\lambda^2 + 2c_v^2)}{(c_u^2\lambda^2 + c_v^2)^{3/2}\lambda^3\sigma_G} h(a_2) \\
&\quad - \left(\frac{c_v^2\mu}{(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G}\right)^2 h(a_2)[a_2 - h(a_2)],
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial(\sigma_G^2)^2} &= \frac{1}{2\sigma_G^4} - \left(\frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G^3} + \frac{|c_v|}{c_u\lambda\sigma_G^2} a_1\right)^2 - \frac{3}{4\sigma_G^4} a_1 h(a_1) \\
&\quad + \frac{1}{4\sigma_G^4} a_1^2 h(a_1)[a_1 - h(a_1)] + \frac{3}{4\sigma_G^4} a_2 h(a_2) \\
&\quad - \frac{1}{4\sigma_G^4} a_2^2 h(a_2)[a_2 - h(a_2)] \\
&= -\left(\frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G^3}\right)^2 - \frac{2|c_v|\mu(c_u^2\lambda^2 + c_v^2)}{c_u^2\lambda^3\sigma_G^5} a_1 - \left(\frac{|c_v|}{c_u\lambda\sigma_G^2}\right)^2 a_1^2 \\
&\quad + \frac{1}{4\sigma_G^4} \left(2 - 3a_1 h(a_1) + a_1^3 h(a_1) - a_1^2 h(a_1)^2 + 3a_2 h(a_2)\right. \\
&\quad \left. - a_2^2 h(a_2)[a_2 - h(a_2)]\right),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \mu} &= -\frac{c_u}{\sigma_G^2}\{1 + h(a_1)[a_1 - h(a_1)]\}\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -\frac{c_u}{\sigma_G^2}\{1 + a_1 h(a_1) - h(a_1)^2\}\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \lambda} &= -\frac{c_u}{|c_v|\sigma_G}\left\{h(a_1) - \left(\frac{2|c_v|\mu}{\lambda\sigma_G} + a_1\right) h(a_1)[a_1 - h(a_1)]\right\}\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= -\frac{c_u}{|c_v|\sigma_G}\left\{h(a_1) - \frac{2|c_v|\mu}{\lambda\sigma_G}[a_1 h(a_1) - h(a_1)^2] - a_1^2 h(a_1)\right. \\
&\qquad \left. + a_1 h(a_1)^2\right\}\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \sigma_G^2} &= \frac{1}{\sigma_G^3}\left\{\frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G} + \frac{|c_v|}{c_u\lambda}a_1 + \frac{c_u\lambda}{2|c_v|}h(a_1)\right. \\
&\qquad \left. - \frac{c_u\lambda}{2|c_v|}a_1 h(a_1)[a_1 - h(a_1)]\right\}\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\sigma_G^3}\left\{\frac{\mu(c_u^2\lambda^2 + c_v^2)}{c_u\lambda^2\sigma_G} + \frac{|c_v|}{c_u\lambda}a_1\right. \\
&\qquad \left. + \frac{c_u\lambda}{2|c_v|}[h(a_1) - a_1^2 h(a_1) + a_1 h(a_1)^2]\right\}\frac{\partial f(\boldsymbol{x}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \mu \partial \lambda} &= -\frac{|c_v|}{\lambda^2\sigma_G}\left(h(a_1) + \left(\frac{2|c_v|\mu}{\lambda\sigma_G} + a_1\right)h(a_1)[a_1 - h(a_1)]\right) \\
&\qquad + \frac{c_v^2}{(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G}\left(h(a_2) - a_2 h(a_2)[a_2 - h(a_2)]\right) \\
&= -\frac{|c_v|}{\lambda^2\sigma_G}\left(h(a_1) + \frac{2|c_v|\mu}{\lambda\sigma_G}[a_1 h(a_1) - h(a_1)^2] + a_1^2 h(a_1) - a_1 h(a_1)^2\right) \\
&\qquad + \frac{c_v^2}{(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G}\left(h(a_2) - a_2 h(a_2)[a_2 - h(a_2)]\right),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \mu \partial \sigma_G^2} &= \frac{\mu(c_u^2\lambda^2 + c_v^2)}{\lambda^2\sigma_G^4} + \frac{|c_v|}{\lambda\sigma_G^3}a_1 - \frac{|c_v|}{2\lambda\sigma_G^3}\left(h(a_1) - a_1 h(a_1)[a_1 - h(a_1)]\right) \\
&\qquad - \frac{1}{2\mu\sigma_G^2}a_2 h(a_2) + \frac{1}{2\mu\sigma_G^2}a_2^2 h(a_2)[a_2 - h(a_2)] \\
&= \frac{\mu(c_u^2\lambda^2 + c_v^2)}{\lambda^2\sigma_G^4} + \frac{|c_v|}{2\lambda\sigma_G^3}\left[2a_1 - h(a_1) + a_1^2 h(a_1) - a_1 h(a_1)^2\right]
\end{aligned}$$

$$-\frac{1}{2\mu\sigma_G^2}a_2h(a_2)+\frac{1}{2\mu\sigma_G^2}a_2^2h(a_2)[a_2-h(a_2)],$$

$$\begin{aligned}\frac{\partial^2\ln f_Y}{\partial\lambda\partial\sigma_G^2} &= \frac{1}{2\lambda\sigma_G^2}\left(\frac{2|c_v|\mu}{\lambda\sigma_G}+a_1\right)\left(h(a_1)-a_1h(a_1)[a_1-h(a_1)]\right)\\ &\quad -\frac{c_v^2\mu}{2(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G^3}\left(h(a_2)-a_2h(a_2)[a_2-h(a_2)]\right)\\ &= \frac{|c_v|\mu}{\lambda^2\sigma_G^3}\left[h(a_1)-a_1^2h(a_1)+a_1h(a_1)^2\right]\\ &\quad +\frac{1}{2\lambda\sigma_G^2}\left[a_1h(a_1)-a_1^3h(a_1)+a_1^2h(a_1)^2\right]\\ &\quad -\frac{c_v^2\mu}{2(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G^3}\left(h(a_2)-a_2h(a_2)[a_2-h(a_2)]\right).\end{aligned}$$

A.2.2 Information matrix in terms of first-order partial derivatives

The form of the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta}=(\boldsymbol{\beta},\mu,\lambda,\sigma_G)$ is given in equation (2.26). Dispensing with the observation subscripts, the components of the information matrix are

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial\ln f_Y}{\partial\boldsymbol{\beta}}\right)^T\right] &= \left\{\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2}\right)^2+\frac{2|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^3}\mathbb{E}[a_1]\right.\\ &\quad +\frac{2\mu(c_u^2\lambda^2+c_v^2)}{|c_v|\lambda\sigma_G^3}\mathbb{E}[h(a_1)]+\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a_1^2]+\frac{2}{\sigma_G^2}\mathbb{E}[a_1h(a_1)]\\ &\quad \left.+\left(\frac{c_u\lambda}{|c_v|\sigma_G}\right)^2\mathbb{E}[h(a_1)^2]\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\mu}\right)^2\right] &= \left(\frac{a_2}{\mu}\right)^2[a_2-h(a_2)]^2\\ &\quad +\frac{2|c_v|}{\mu\lambda\sigma_G}a_2[a_2-h(a_2)]\left(\mathbb{E}[a_1]-\mathbb{E}[h(a_1)]\right)\end{aligned}$$

$$+\left(\frac{|c_v|}{\lambda\sigma_G}\right)^2\left(\mathbb{E}[a_1^2]-2\mathbb{E}[a_1h(a_1)]+\mathbb{E}[h(a_1)^2]\right),$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial\lambda}\right)^2\right] &= \left(\frac{2|c_v|\mu}{\lambda^2\sigma_G}\right)^2\mathbb{E}[h(a_1)^2]+\frac{4|c_v|\mu}{\lambda^3\sigma_G}\mathbb{E}[a_1h(a_1)^2]\\ &-\frac{4|c_v|^3\mu^2}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^4\sigma_G^2}h(a_2)\mathbb{E}[h(a_1)]+\frac{1}{\lambda^2}\mathbb{E}[a_1^2h(a_1)^2]\\ &-\frac{2c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^3\sigma_G}h(a_2)\mathbb{E}[a_1h(a_1)]\\ &+\left(\frac{c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}\right)^2h(a_2)^2,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial\sigma_G^2}\right)^2\right] &= \frac{1}{4\sigma_G^4}\left[-\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G}\right)^2+a_2h(a_2)+1\right]^2\\ &-\frac{1}{\sigma_G^2}\left[-\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G}\right)^2+a_2h(a_2)+1\right]\times\\ &\left(\frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^3}\mathbb{E}[a_1]+\frac{1}{2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a_1^2]+\frac{1}{2\sigma_G^2}\mathbb{E}[a_1h(a_1)]\right)\\ &+\left(\frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^3}\right)^2\mathbb{E}[a_1^2]+\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^3\mathbb{E}[a_1^3]\\ &+\frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^5}\mathbb{E}[a_1^2h(a_1)]+\frac{1}{4}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^4\mathbb{E}[a_1^4]\\ &+\frac{1}{2\sigma_G^2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a_1^3h(a_1)]+\frac{1}{4\sigma_G^4}\mathbb{E}[a_1^2h(a_1)^2],\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial\mu}\right)\right] &= \left\{\left[\frac{\mu(c_u^2\lambda^2+c_v^2)}{\lambda^2\sigma_G^2}-\frac{a_2}{\mu}h(a_2)\right]\times\right.\\ &\left.\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2}+\frac{|c_v|}{c_u\lambda\sigma_G}\mathbb{E}[a_1]+\frac{c_u\lambda}{|c_v|\sigma_G}\mathbb{E}[h(a_1)]\right)\right.\end{aligned}$$

$$
\begin{aligned}
&+\frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^3\sigma_G^3}\left(\mathbb{E}[a_1]-\mathbb{E}[h(a_1)]\right)\\
&+\frac{1}{c_u}\left(\frac{|c_v|}{\lambda\sigma_G}\right)^2\left(\mathbb{E}[a_1^2]-\mathbb{E}[a_1h(a_1)]\right)\\
&+\frac{c_u}{\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)-h(a_1)^2]\right)\Bigg\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial\ln f_Y}{\partial\lambda}\right)\right] \quad = \quad &\left\{\frac{2|c_v|\mu^2(c_u^2\lambda^2+c_v^2)}{c_u\lambda^4\sigma_G^3}\mathbb{E}[h(a_1)]\right.\\
&+\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^3\sigma_G^2}\mathbb{E}[a_1h(a_1)]-\frac{c_v^2\mu^2(c_u^2\lambda^2+c_v^2)^{1/2}}{c_u\lambda^4\sigma_G^3}h(a_2)\\
&+\frac{2c_v^2\mu}{c_u\lambda^3\sigma_G^2}\mathbb{E}[a_1h(a_1)]+\frac{|c_v|}{c_u\lambda^2\sigma_G}\mathbb{E}[a_1^2h(a_1)]\\
&-\frac{|c_v|^3\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}c_u\lambda^3\sigma_G^2}h_2(a_2)\mathbb{E}[a_1]+\frac{2c_u\mu}{\lambda\sigma_G^2}\mathbb{E}[h(a_1)^2]\\
&\left.+\frac{c_u}{|c_v|\sigma_G}\mathbb{E}[a_1h(a_1)^2]-\frac{c_u|c_v|\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda\sigma_G^2}h_2(a_2)\mathbb{E}[h(a_1)]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial\ln f_Y}{\partial\sigma_G^2}\right)\right] \quad = \quad &\\
&-\left\{-\frac{1}{2\sigma_G^2}\left[-\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G}\right)^2+a_2h(a_2)+1\right]\times\right.\\
&\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2}+\frac{|c_v|}{c_u\lambda\sigma_G}\mathbb{E}[a_1]+\frac{c_u\lambda}{|c_v|\sigma_G}\mathbb{E}[h(a_1)]\right)\\
&+\frac{|c_v|\mu^2(c_u^2\lambda^2+c_v^2)^2}{c_u^3\lambda^5\sigma_G^5}\mathbb{E}[a_1]+\frac{3c_v^2\mu(c_u^2\lambda^2+c_v^2)}{2c_u^3\lambda^4\sigma_G^4}\mathbb{E}[a_1^2]\\
&+\frac{3\mu(c_u^2\lambda^2+c_v^2)}{2c_u\lambda^2\sigma_G^4}\mathbb{E}[a_1h(a_1)]+\frac{1}{2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^3\mathbb{E}[a_1^3]\\
&\left.+\frac{|c_v|}{c_u\lambda\sigma_G^3}\mathbb{E}[a_1^2h(a_1)]+\frac{c_u\lambda}{2|c_v|\sigma_G^3}\mathbb{E}[a_1h(a_1)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}
$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \mu}\right)\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\right] &= \left[\frac{\mu(c_u^2\lambda^2+c_v^2)}{\lambda^2\sigma_G^2} - \frac{a_2}{\mu}h(a_2)\right]\times \\
&\left(\frac{2|c_v|\mu}{\lambda^2\sigma_G}\mathbb{E}[h(a_1)] + \frac{1}{\lambda}\mathbb{E}[a_1h(a_1)] - \frac{c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}h(a_2)\right) \\
&+\frac{2c_v^2\mu}{\lambda^3\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)] - \mathbb{E}[h(a_1)^2]\right) + \frac{|c_v|}{\lambda^2\sigma_G}\left(\mathbb{E}[a_1^2h(a_1)] - \mathbb{E}[a_1h(a_1)^2]\right) \\
&-\frac{|c_v|^3\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^3\sigma_G^2}h(a_2)\left(\mathbb{E}[a_1] - \mathbb{E}[h(a_1)]\right),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \mu}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] &= \left[\frac{\mu(c_u^2\lambda^2+c_v^2)}{\lambda^2\sigma_G^2} - \frac{a_2}{\mu}h(a_2)\right]\times \\
&\left(\frac{1}{2\sigma_G^2}\left[-\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G}\right)^2 + a_2h(a_2) + 1\right] - \frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^3}\mathbb{E}[a_1]\right. \\
&\left.-\frac{1}{2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a_1^2] - \frac{1}{2\sigma_G^2}\mathbb{E}[a_1h(a_1)]\right) \\
&+\frac{|c_v|}{2\lambda\sigma_G^3}\left[-\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G}\right)^2 + a_2h(a_2) + 1\right]\left(\mathbb{E}[a_1] - \mathbb{E}[h(a_1)]\right) \\
&-\frac{c_v^2\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^4\sigma_G^4}\left(\mathbb{E}[a_1^2] - \mathbb{E}[a_1h(a_1)]\right) \\
&-\frac{1}{2c_u^2}\left(\frac{|c_v|}{\lambda\sigma_G}\right)^3\left(\mathbb{E}[a_1^3] - \mathbb{E}[a_1^2h(a_1)]\right) \\
&-\frac{|c_v|}{2\lambda\sigma_G^3}\left(\mathbb{E}[a_1^2h(a_1)] - \mathbb{E}[a_1h(a_1)^2]\right),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] &= \frac{1}{2\sigma_G^2}\left[-\left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G}\right)^2 + a_2h(a_2) + 1\right]\times \\
&\left(\frac{2|c_v|\mu}{\lambda^2\sigma_G}\mathbb{E}[h(a_1)] + \frac{1}{\lambda}\mathbb{E}[a_1h(a_1)] - \frac{c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}h(a_2)\right) \\
&-\frac{2c_v^2\mu^2(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^5\sigma_G^4}\mathbb{E}[a_1h(a_1)] - \frac{\mu}{c_u^2\lambda}\left(\frac{|c_v|}{\lambda\sigma_G}\right)^3\mathbb{E}[a_1^2h(a_1)] \\
&-\frac{|c_v|\mu}{\lambda^2\sigma_G^3}\mathbb{E}[a_1h(a_1)^2] - \frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^4\sigma_G^3}\mathbb{E}[a_1^2h(a_1)]
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2\lambda}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a_1^3h(a_1)] - \frac{1}{2\lambda\sigma_G^2}\mathbb{E}[a_1^2h(a_1)^2] \\
&+\frac{c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}h(a_2)\times \\
&\left(\frac{|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^3}\mathbb{E}[a_1] + \frac{1}{2}\left(\frac{|c_v|}{c_u\lambda\sigma_G}\right)^2\mathbb{E}[a_1^2] + \frac{1}{2\sigma_G^2}\mathbb{E}[a_1h(a_1)]\right).
\end{aligned}
$$

A.2.3 Information matrix in terms of second-order partial derivatives

Equation (2.27) gives an alternative formulation for the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \lambda, \sigma_G)$. The components of the per observation expected Fisher information matrix are

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] &= \\
&-\left\{-\frac{1}{\sigma_G^2}+\left(\frac{c_u\lambda}{|c_v|\sigma_G}\right)^2\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right)\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^T}\right) \\
&+\left\{\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^2}+\frac{|c_v|}{c_u\lambda\sigma_G}\mathbb{E}[a_1]+\frac{c_u\lambda}{|c_v|\sigma_G}\mathbb{E}[h(a_1)]\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\mu^2}\right] &= \left(\frac{c_u}{\sigma_G}\right)^2-\left(\frac{|c_v|}{\lambda\sigma_G}\right)^2\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right) \\
&\quad+\left(\frac{a_2}{\mu}\right)^2 h(a_2)[a_2-h(a_2)],
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\lambda^2}\right] &= -\frac{2|c_v|\mu}{\lambda^3\sigma_G}\mathbb{E}[h(a_1)]-\left(\frac{2|c_v|\mu}{\lambda^2\sigma_G}\right)^2\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right) \\
&\quad-\frac{4|c_v|\mu}{\lambda^3\sigma_G}\left(\mathbb{E}[a_1^2h(a_1)]-\mathbb{E}[a_1h(a_1)^2]\right) \\
&\quad-\frac{1}{\lambda^2}\left(\mathbb{E}[a_1^3h(a_1)]-\mathbb{E}[a_1^2h(a_1)^2]\right)+\frac{c_v^2\mu(3c_u^2\lambda^2+2c_v^2)}{(c_u^2\lambda^2+c_v^2)^{3/2}\lambda^3\sigma_G}h(a_2) \\
&\quad+\left(\frac{c_v^2\mu}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}\right)^2 h(a_2)[a_2-h(a_2)],
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(\sigma_G^2)^2}\right] &= \left(\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G^3}\right)^2 + \frac{2|c_v|\mu(c_u^2\lambda^2+c_v^2)}{c_u^2\lambda^3\sigma_G^5}\mathbb{E}[a_1] + \left(\frac{|c_v|}{c_u\lambda\sigma_G^2}\right)^2\mathbb{E}[a_1^2] \\
&\quad -\frac{1}{4\sigma_G^4}\left(2 - 3\mathbb{E}[a_1h(a_1)] + \mathbb{E}[a_1^3h(a_1)] - \mathbb{E}[a_1^2h(a_1)^2]\right. \\
&\quad \left.+3a_2h(a_2) - a_2^2h(a_2)[a_2 - h(a_2)]\right),
\end{aligned}
$$

$$
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\mu}\right] = \frac{c_u}{\sigma_G^2}\{1 + \mathbb{E}[a_1h(a_1)] - \mathbb{E}[h(a_1)^2]\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\lambda}\right] &= \frac{c_u}{|c_v|\sigma_G}\left\{\mathbb{E}[h(a_1)] - \frac{2|c_v|\mu}{\lambda\sigma_G}\left(\mathbb{E}[a_1h(a_1)] - \mathbb{E}[h(a_1)^2]\right) - \mathbb{E}[a_1^2h(a_1)]\right. \\
&\quad \left.+\mathbb{E}[a_1h(a_1)^2]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_G^2}\right] &= -\frac{1}{\sigma_G^3}\left\{\frac{\mu(c_u^2\lambda^2+c_v^2)}{c_u\lambda^2\sigma_G} + \frac{|c_v|}{c_u\lambda}\mathbb{E}[a_1]\right. \\
&\quad \left.+\frac{c_u\lambda}{2|c_v|}\left(\mathbb{E}[h(a_1)] - \mathbb{E}[a_1^2h(a_1)] + \mathbb{E}[a_1h(a_1)^2]\right)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\mu\partial\lambda}\right] &= \\
&\frac{|c_v|}{\lambda^2\sigma_G}\left(\mathbb{E}[h(a_1)] + \frac{2|c_v|\mu}{\lambda\sigma_G}\left(\mathbb{E}[a_1h(a_1)] - \mathbb{E}[h(a_1)^2]\right) + \mathbb{E}[a_1^2h(a_1)] - \mathbb{E}[a_1h(a_1)^2]\right) \\
&-\frac{c_v^2}{(c_u^2\lambda^2+c_v^2)^{1/2}\lambda^2\sigma_G}\left(h(a_2) - a_2h(a_2)[a_2 - h(a_2)]\right),
\end{aligned}
$$

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\mu\partial\sigma_G^2}\right] &= -\frac{\mu(c_u^2\lambda^2+c_v^2)}{\lambda^2\sigma_G^4} \\
&\quad -\frac{|c_v|}{2\lambda\sigma_G^3}\left(2\mathbb{E}[a_1] - \mathbb{E}[h(a_1)] + \mathbb{E}[a_1^2h(a_1)] - \mathbb{E}[a_1h(a_1)^2]\right) \\
&\quad +\frac{1}{2\mu\sigma_G^2}a_2h(a_2) - \frac{1}{2\mu\sigma_G^2}a_2^2h(a_2)[a_2 - h(a_2)],
\end{aligned}
$$

$$
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\lambda\partial\sigma_G^2}\right] = -\frac{|c_v|\mu}{\lambda^2\sigma_G^3}\left(\mathbb{E}[h(a_1)] - \mathbb{E}[a_1^2h(a_1)] + \mathbb{E}[a_1h(a_1)^2]\right)
$$

$$-\frac{1}{2\lambda\sigma_G^2}\left(\mathbb{E}[a_1 h(a_1)] - \mathbb{E}[a_1^3 h(a_1)] + \mathbb{E}[a_1^2 h(a_1)^2]\right)$$
$$+\frac{c_v^2\mu}{2(c_u^2\lambda^2 + c_v^2)^{1/2}\lambda^2\sigma_G^3}\left(h(a_2) - a_2 h(a_2)[a_2 - h(a_2)]\right).$$

A.3 Calculations for the Normal-Gamma Model

In this section, the detailed calculations for deriving the information matrix for the normal-gamma model in Section 2.5 of Chapter 2 are given.

The probability density function of $U \sim \text{Gamma}(\alpha, \sigma_u)$ is

$$f_U(u;\alpha,\sigma_u) = \frac{u^{\alpha-1}}{\Gamma(\alpha)\sigma_u^\alpha}\exp\left\{-\frac{u}{\sigma_u}\right\}, \qquad u \geq 0,\ \alpha,\sigma_u > 0, \tag{A.7}$$

with mean and variance

$$\begin{aligned}\mathbb{E}(U) &= \alpha\sigma_u,\\ Var(U) &= \alpha\sigma_u^2.\end{aligned}$$

The density of V is given in equation (2.4) with mean and variance given in equation (2.5). The joint probability density function of U and V is

$$\begin{aligned}f_{U,V}(u,v) &= f_U(u)\cdot f_V(v)\\ &= \frac{u^{\alpha-1}}{\Gamma(\alpha)\sigma_u^\alpha\sqrt{2\pi}\sigma_v}\exp\left\{-\frac{u}{\sigma_u} - \frac{v^2}{2\sigma_v^2}\right\}.\end{aligned} \tag{A.8}$$

The joint density function of U and $E = c_u U + c_v V$ can be derived using equation (C.1) in Appendix C and is given by

$$\begin{aligned}f_{U,E}(u,\varepsilon) &= \frac{1}{|c_v|}f_{U,V}\left(u, \frac{\varepsilon - c_u u}{c_v}\right)\\ &= \frac{u^{\alpha-1}}{|c_v|\Gamma(\alpha)\sigma_u^\alpha\sqrt{2\pi}\sigma_v}\exp\left\{-\frac{u}{\sigma_u} - \frac{(\varepsilon - c_u u)^2}{2c_v^2\sigma_v^2}\right\}\\ &= \frac{u^{\alpha-1}}{|c_v|\Gamma(\alpha)\sigma_u^\alpha\sqrt{2\pi}\sigma_v}\exp\left\{-\frac{1}{2}\left[\frac{c_u^2}{c_v^2\sigma_v^2}u^2 - 2\left(\frac{c_u\varepsilon}{c_v^2\sigma_v^2} - \frac{1}{\sigma_u}\right)u + \frac{\varepsilon^2}{c_v^2\sigma_v^2}\right]\right\}.\end{aligned} \tag{A.9}$$

If we let $K = \dfrac{1}{|c_v|\Gamma(\alpha)\sigma_u^\alpha\sqrt{2\pi}\sigma_v}$, $A = \dfrac{c_u^2}{c_v^2\sigma_v^2}$, $B = \dfrac{c_u\varepsilon}{c_v^2\sigma_v^2} - \dfrac{1}{\sigma_u}$ and $C = \dfrac{\varepsilon^2}{c_v^2\sigma_v^2}$ then the joint density function of U and E becomes

$$f_{U,E}(u,\varepsilon) = u^{\alpha-1}K\exp\left\{-\frac{1}{2}[Au^2 - 2Bu + C]\right\}.$$

When the joint density of U and E is of this form, the marginal density of E is given by equations (C.4) and (C.5) in Appendix C as

$$\begin{aligned} f_E(\varepsilon) &= K\exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\}\int_0^\infty u^{\alpha-1}\exp\left\{-\frac{1}{2}\left(\frac{u - B/A}{1/\sqrt{A}}\right)^2\right\}du \\ &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\}\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u - B/A}{1/\sqrt{A}}\right)du \\ &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\}\Phi\left(\frac{B}{\sqrt{A}}\right)\mathbb{E}[Q^{\alpha-1}], \end{aligned}$$

where random variable Q has a normal distribution, with mean B/A and variance $1/A$, which is truncated from below at zero, i.e. $Q \sim N^+\left(\dfrac{B}{A}, \dfrac{1}{A}\right)$ with

$$\begin{aligned} \frac{B}{A} &= \frac{\varepsilon}{c_u} - \frac{c_v^2\sigma_v^2}{c_u^2\sigma_u}, \\ \frac{1}{A} &= \frac{c_v^2\sigma_v^2}{c_u^2}. \end{aligned}$$

$\mathbb{E}[Q^{\alpha-1}]$ is a fractional moment of the nonnegative truncated normal distribution of Q. Appendix C.5 provides further details on truncated normal distributions. The components of the marginal density of E are

$$\begin{aligned} K\sqrt{\frac{2\pi}{A}} &= \frac{1}{|c_u|\Gamma(\alpha)\sigma_u^\alpha}, \\ C - \frac{B^2}{A} &= \frac{2\varepsilon}{c_u\sigma_u} - \frac{c_v^2\sigma_v^2}{c_u^2\sigma_u^2}, \\ \frac{B}{\sqrt{A}} &= \frac{c_u\varepsilon}{|c_uc_v|\sigma_v} - \frac{|c_v|\sigma_v}{|c_u|\sigma_u}, \end{aligned}$$

thus the marginal density of E is given by

$$\begin{aligned} f_E(\varepsilon) &= \frac{1}{|c_u|\Gamma(\alpha)\sigma_u^\alpha}\exp\left\{-\frac{\varepsilon}{c_u\sigma_u}+\frac{c_v^2\sigma_v^2}{2c_u^2\sigma_u^2}\right\}\times \\ &\qquad \int_0^\infty u^{\alpha-1}\frac{|c_u|}{|c_v|\sigma_v}\phi\left(\frac{|c_u|}{|c_v|\sigma_v}u-\frac{c_u\varepsilon}{|c_uc_v|\sigma_v}+\frac{|c_v|\sigma_v}{|c_u|\sigma_u}\right)du \\ &= \frac{1}{|c_u|\Gamma(\alpha)\sigma_u^\alpha}\exp\left\{-\frac{\varepsilon}{c_u\sigma_u}+\frac{c_v^2\sigma_v^2}{2c_u^2\sigma_u^2}\right\}\Phi\left(\frac{c_u\varepsilon}{|c_uc_v|\sigma_v}-\frac{|c_v|\sigma_v}{|c_u|\sigma_u}\right)\mathbb{E}[Q^{\alpha-1}], \end{aligned}$$

with mean and variance that can be derived using equations (C.7) and (C.8) in Appendix C and which are given by

$$\begin{aligned} \mathbb{E}[E] &= \tilde{c_u}\sigma_u, \\ Var(E) &= \tilde{\tilde{c_u}}^2\sigma_u^2+c_v^2\sigma_v^2, \end{aligned}$$

where $\tilde{c_u} = c_u\alpha$ and $\tilde{\tilde{c_u}}^2 = c_u^2\alpha$.

A.3.1 Log-likelihood function

The log-likelihood function of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$ is given in equation (2.31). Because the log-likelihood function under the normal-gamma specification contains an integral, calculation of the derivatives of $\ln f_Y(y;\boldsymbol{\theta})$ require slightly more working than under the alternative specifications of previous sections. Detailed calculations for the derivatives of the the integral appearing in the log-likelihood function have not been shown here but can found in Appendix A.4.

Using the reparameterisation $\varepsilon = y - f(\boldsymbol{x},\boldsymbol{\beta})$, the first-order derivatives of $\ln f_Y(y;\boldsymbol{\theta})$ with respect to the parameters of interest are

$$\begin{aligned} \frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= \frac{1}{c_v^2\sigma_v^2}\left\{y - f(\boldsymbol{x},\boldsymbol{\beta}) - c_u\mathbb{E}[U|E]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \frac{1}{c_v^2\sigma_v^2}\left\{\varepsilon - c_u\mathbb{E}[U|E]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \alpha} &= -\psi(\alpha) + \ln\left(\frac{1}{\sigma_u}\right) + \mathbb{E}[\ln U|E], \\
\frac{\partial \ln f_Y}{\partial (1/\sigma_u)} &= \alpha\sigma_u - \mathbb{E}[U|E], \\
\frac{\partial \ln f_Y}{\partial \sigma_v^2} &= -\frac{1}{2\sigma_v^2} + \frac{1}{2c_v^2\sigma_v^4} \times \\
&\qquad \left([y - f(\boldsymbol{x},\boldsymbol{\beta})]^2 + c_u^2\mathbb{E}[U^2|E] - 2c_u[y - f(\boldsymbol{x},\boldsymbol{\beta})]\mathbb{E}[U|E]\right) \\
&= -\frac{1}{2\sigma_v^2} + \frac{1}{2c_v^2\sigma_v^4}\left(\varepsilon^2 + c_u^2\mathbb{E}[U^2|E] - 2c_u\varepsilon\mathbb{E}[U|E]\right),
\end{aligned}$$

where

$$\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{d\ln\Gamma(\alpha)}{d\alpha},$$

is the digamma function and equation (C.11) in Appendix C gives

$$\mathbb{E}[g(U)|E] = \frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}.$$

The corresponding second-order derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \frac{1}{c_v^2\sigma_v^2}\left\{-1 + \frac{c_u^2}{c_v^2\sigma_v^2}Var(U|E)\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T \\
&\qquad + \frac{1}{c_v^2\sigma_v^2}\left\{y - f(\boldsymbol{x},\boldsymbol{\beta}) - c_u\mathbb{E}[U|E]\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \\
&= \frac{1}{c_v^2\sigma_v^2}\left\{-1 + \frac{c_u^2}{c_v^2\sigma_v^2}Var(U|E)\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\right)^T \\
&\qquad + \frac{1}{c_v^2\sigma_v^2}\left\{\varepsilon - c_u\mathbb{E}[U|E]\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}, \\
\frac{\partial^2 \ln f_Y}{\partial\alpha^2} &= -\psi_1(\alpha) + Var(\ln U|E),
\end{aligned}$$

$$\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)^2} = -\alpha\sigma_u^2 + Var(U|E),$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial(\sigma_v^2)^2} &= \frac{1}{2\sigma_v^4} - \frac{1}{c_v^2\sigma_v^6} \times \\ &\quad \left([y - f(\boldsymbol{x},\boldsymbol{\beta})]^2 + c_u^2\mathbb{E}[U^2|E] - 2c_u[y - f(\boldsymbol{x},\boldsymbol{\beta})]\mathbb{E}[U|E]\right) \\ &\quad + \left(\frac{c_u}{c_v^2\sigma_v^4}\right)^2 \left[\frac{c_u^2}{4}Var(U^2|E) - c_u[y - f(\boldsymbol{x},\boldsymbol{\beta})]Cov(U,U^2|E)\right. \\ &\quad \left. + [y - f(\boldsymbol{x},\boldsymbol{\beta})]^2 Var(U|E)\right] \\ &= \frac{1}{2\sigma_v^4} - \frac{1}{c_v^2\sigma_v^6}\left(\varepsilon^2 + c_u^2\mathbb{E}[U^2|E] - 2c_u\varepsilon\mathbb{E}[U|E]\right) + \left(\frac{c_u}{c_v^2\sigma_v^4}\right)^2 \times \\ &\quad \left[\frac{c_u^2}{4}Var(U^2|E) - c_u\varepsilon Cov(U,U^2|E) + \varepsilon^2 Var(U|E)\right],\end{aligned}$$

$$\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\alpha} = -\frac{c_u}{c_v^2\sigma_v^2}Cov(U,\ln U|E)\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},$$

$$\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)} = \frac{c_u}{c_v^2\sigma_v^2}Var(U|E)\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_v^2} &= -\frac{1}{c_v^2\sigma_v^4}\left\{y - f(\boldsymbol{x},\boldsymbol{\beta}) - c_u\mathbb{E}[U|E] + \frac{c_u^3}{2c_v^2\sigma_v^2}Cov(U,U^2|E)\right. \\ &\quad \left. - \frac{c_u^2}{c_v^2\sigma_v^2}[y - f(\boldsymbol{x},\boldsymbol{\beta})]Var(U|E)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \\ &= -\frac{1}{c_v^2\sigma_v^4}\left\{\varepsilon - c_u\mathbb{E}[U|E] + \frac{c_u^3}{2c_v^2\sigma_v^2}Cov(U,U^2|E)\right. \\ &\quad \left. - \frac{c_u^2}{c_v^2\sigma_v^2}\varepsilon Var(U|E)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},\end{aligned}$$

$$\frac{\partial^2 \ln f_Y}{\partial\alpha\partial(1/\sigma_u)} = \sigma_u - Cov(U,\ln U|E),$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial\alpha\partial\sigma_v^2} &= \frac{c_u^2}{2c_v^2\sigma_v^4}Cov(U^2,\ln U|E) - \frac{c_u}{c_v^2\sigma_v^4}[y - f(\boldsymbol{x},\boldsymbol{\beta})]Cov(U,\ln U|E) \\ &= \frac{c_u^2}{2c_v^2\sigma_v^4}Cov(U^2,\ln U|E) - \frac{c_u}{c_v^2\sigma_v^4}\varepsilon Cov(U,\ln U|E),\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)\partial\sigma_v^2} &= -\frac{c_u^2}{2c_v^2\sigma_v^4}Cov(U,U^2|E) + \frac{c_u}{c_v^2\sigma_v^4}[y - f(\boldsymbol{x},\boldsymbol{\beta})]Var(U|E) \\ &= -\frac{c_u^2}{2c_v^2\sigma_v^4}Cov(U,U^2|E) + \frac{c_u}{c_v^2\sigma_v^4}\varepsilon Var(U|E),\end{aligned}$$

where

$$\psi_1(\alpha) = \frac{d\psi(\alpha)}{d\alpha} = \frac{d^2 \ln \Gamma(\alpha)}{d\alpha^2},$$

is the trigamma function.

A.3.2 Information matrix in terms of first-order partial derivatives

The form of the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$ is given in equation (2.33). Dispensing with the observation subscripts, the components of the information matrix are

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)^T\right] &= \left(\frac{1}{c_v^2\sigma_v^2}\right)^2 \left\{\mathbb{E}[E^2] - 2c_u\mathbb{E}(E\cdot\mathbb{E}[U|E]) \right. \\ &\quad \left. + c_u^2\mathbb{E}(\mathbb{E}[U|E]^2)\right\}\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \alpha}\right)^2\right] &= \left[-\psi(\alpha) + \ln\left(\frac{1}{\sigma_u}\right)\right]^2 \\ &\quad + 2\left[-\psi(\alpha) + \ln\left(\frac{1}{\sigma_u}\right)\right]\mathbb{E}\left(\mathbb{E}[\ln U|E]\right) + \mathbb{E}\left(\mathbb{E}[\ln U|E]^2\right),\end{aligned}$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial(1/\sigma_u)}\right)^2\right] = \alpha^2\sigma_u^2 - 2\alpha\sigma_u\mathbb{E}\left(\mathbb{E}[U|E]\right) + \mathbb{E}\left(\mathbb{E}[U|E]^2\right),$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)^2\right] \quad = \quad & \left(\frac{1}{2\sigma_v^2}\right)^2 - \frac{1}{2c_v^2\sigma_v^6}\left\{\mathbb{E}[E^2] + c_u^2\mathbb{E}\left(\mathbb{E}[U^2|E]\right) - 2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\} \\ & + \left(\frac{1}{2c_v^2\sigma_v^4}\right)^2\left\{\mathbb{E}[E^4] + 2c_u^2\mathbb{E}\left(E^2\cdot\mathbb{E}[U^2|E]\right) - 4c_u\mathbb{E}\left(E^3\cdot\mathbb{E}[U|E]\right)\right. \\ & + c_u^4\mathbb{E}\left(\mathbb{E}[U^2|E]^2\right) - 2c_u^3\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\cdot\mathbb{E}[U^2|E]\right) \\ & \left. + 4c_u^2\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]^2\right)\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \alpha}\right)\right] \quad = \quad & \frac{1}{c_v^2\sigma_v^2}\times \\ & \left\{\left[-\psi(\alpha) + \ln\left(\frac{1}{\sigma_u}\right)\right]\left(\mathbb{E}[E] - c_u\mathbb{E}[U|E]\right)\right. \\ & \left. + \mathbb{E}\left(E\cdot\mathbb{E}[\ln U|E]\right) + c_u\mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[\ln U|E]\right)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)\right] \quad = \quad & \frac{1}{c_v^2\sigma_v^2}\times \\ & \left\{\alpha\sigma_u\mathbb{E}[E] - c_u\alpha\sigma_u\mathbb{E}[U|E] - \mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right. \\ & \left. + c_u\mathbb{E}\left(\mathbb{E}[U|E]^2\right)\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] \quad = \quad & -\frac{1}{2c_v^2\sigma_v^4}\left\{\mathbb{E}[E] - c_u\mathbb{E}\left(\mathbb{E}[U|E]\right)\right\} \\ & + \frac{1}{2c_v^4\sigma_v^6}\left\{\mathbb{E}[E^3] + c_u^2\mathbb{E}\left(E\cdot\mathbb{E}[U^2|E]\right) - 2c_u\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]\right)\right\} \\ & - \frac{c_u}{2c_v^4\sigma_v^6}\left\{\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]\right) + c_u^2\mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[U^2|E]\right)\right. \\ & \left. - 2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]^2\right)\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \alpha}\right)\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)\right] &= \left[-\psi(\alpha)+\ln\left(\frac{1}{\sigma_u}\right)\right]\{\alpha\sigma_u - \mathbb{E}\left(\mathbb{E}[U|E]\right)\} \\ &+\alpha\sigma_u\mathbb{E}\left(\mathbb{E}[\ln U|E]\right) - \mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[\ln U|E]\right),\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \alpha}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] &= -\frac{1}{2\sigma_v^2}\left\{-\psi(\alpha)+\ln\left(\frac{1}{\sigma_u}\right)+\mathbb{E}\left(\mathbb{E}[\ln U|E]\right)\right\} \\ &+\frac{1}{2c_v^2\sigma_v^4}\left[-\psi(\alpha)+\ln\left(\frac{1}{\sigma_u}\right)\right]\times \\ &\left\{\mathbb{E}[E^2]+c_u^2\mathbb{E}\left(\mathbb{E}[U^2|E]\right)-2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\} \\ &+\frac{1}{2c_v^2\sigma_v^4}\left\{\mathbb{E}\left(E^2\cdot\mathbb{E}[\ln U|E]\right)+c_u^2\mathbb{E}\left(\mathbb{E}[U^2|E]\cdot\mathbb{E}[\ln U|E]\right)\right. \\ &\left.-2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\cdot\mathbb{E}[\ln U|E]\right)\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] &= -\frac{1}{2\sigma_v^2}\left\{\alpha\sigma_u-\mathbb{E}\left(\mathbb{E}[U|E]\right)\right\} \\ &+\frac{\alpha\sigma_u}{2c_v^2\sigma_v^4}\left\{\mathbb{E}[E^2]+c_u^2\mathbb{E}\left(\mathbb{E}[U^2|E]\right)-2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\} \\ &-\frac{1}{2c_v^2\sigma_v^4}\left\{\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]\right)+c_u^2\mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[U^2|E]\right)\right. \\ &\left.-2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]^2\right)\right\}.\end{aligned}$$

A.3.3 Information matrix in terms of second-order partial derivatives

An alternative formulation for the partitioned per observation expected Fisher information matrix of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$ is given in equation (2.32). The components of this per observation expected Fisher information matrix are

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = -\frac{1}{c_v^2\sigma_v^2}\left\{-1+\frac{c_u^2}{c_v^2\sigma_v^2}\mathbb{E}\left[Var(U|E)\right]\right\}\times$$

$$\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)$$
$$-\frac{1}{c_v^2\sigma_v^2}\left\{\mathbb{E}[E]-c_u\mathbb{E}\left(\mathbb{E}[U|E]\right)\right\}\frac{\partial^2 f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T},$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\alpha^2}\right] = \psi_1(\alpha)-\mathbb{E}\left[Var(\ln U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)^2}\right] = \alpha\sigma_u^2-\mathbb{E}\left[Var(U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(\sigma_v^2)^2}\right] = -\frac{1}{2\sigma_v^4}+\frac{1}{c_v^2\sigma_v^6}\times \left\{\mathbb{E}[E^2]+c_u^2\mathbb{E}\left(\mathbb{E}[U^2|E]\right)-2c_u\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\} -\left(\frac{c_u}{c_v^2\sigma_v^4}\right)^2\times \left\{\frac{c_u^2}{4}\mathbb{E}\left[Var(U^2|E)\right]-c_u\mathbb{E}\left[E\cdot Cov(U,U^2|E)\right] +\mathbb{E}\left[E^2\cdot Var(U|E)\right]\right\},$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\alpha}\right] = \frac{c_u}{c_v^2\sigma_v^2}\mathbb{E}\left[Cov(U,\ln U|E)\right]\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)}\right] = -\frac{c_u}{c_v^2\sigma_v^2}\mathbb{E}\left[Var(U|E)\right]\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_v^2}\right] = \frac{1}{c_v^2\sigma_v^4}\left\{\mathbb{E}[E]-c_u\mathbb{E}\left(\mathbb{E}[U|E]\right)+\frac{c_u^3}{2c_v^2\sigma_v^2}\mathbb{E}\left[Cov(U,U^2|E)\right] -\frac{c_u^2}{c_v^2\sigma_v^2}\mathbb{E}\left[E\cdot Var(U|E)\right]\right\}\frac{\partial f(\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}},$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\alpha\partial(1/\sigma_u)}\right] = -\sigma_u+\mathbb{E}\left[Cov(U,\ln U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\alpha\partial\sigma_v^2}\right] = -\frac{c_u^2}{2c_v^2\sigma_v^4}\mathbb{E}\left[Cov(U^2,\ln U|E)\right]+\frac{c_u}{c_v^2\sigma_v^4}\mathbb{E}\left[E\cdot Cov(U,\ln U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)\partial\sigma_v^2}\right] = \frac{c_u^2}{2c_v^2\sigma_v^4}\mathbb{E}\left[Cov(U,U^2|E)\right] - \frac{c_u}{c_v^2\sigma_v^4}\mathbb{E}\left[E\cdot Var(U|E)\right].$$

A.4 Further Calculations for the Normal-Gamma Model

Let

$$a = \frac{u - B/A}{1/\sqrt{A}} = -\frac{|c_u|}{|c_v|\sigma_v}u + \frac{c_u\varepsilon}{|c_uc_v|\sigma_v} - \frac{|c_v|\sigma_v}{|c_u|\sigma_u},$$

and let $f(u)$ be a function such that

$$f(u) = \int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du,$$

where $g(u)$ is a function of u, and possibly α. The derivative of $\ln f(u)$ with respect to $\boldsymbol{\theta}$ is given by

$$\frac{\partial \ln f(u)}{\partial\boldsymbol{\theta}} = \frac{1}{f(u)}\cdot\frac{\partial f(u)}{\partial\boldsymbol{\theta}}.$$

The derivative of $\ln f(u)$ for $g(u) = 1$ appears in the first-order derivatives of $\ln f_Y(y;\boldsymbol{\theta})$ under the normal-gamma specification in Section 2.5 of Chapter 2. Because the limits of integration do not depend on the parameters, differentiation can be taken inside the integral. The derivatives of $f(u)$ with respect to the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \alpha, \sigma_u, \sigma_v)$ are

$$\begin{aligned}\frac{\partial f(u)}{\partial\boldsymbol{\beta}} &= \int_0^\infty g(u)u^{\alpha-1}\left(-\frac{c_u}{c_v^2\sigma_v^2}u + \frac{\varepsilon}{c_v^2\sigma_v^2} - \frac{1}{c_u\sigma_u}\right)\phi(-a)\,du \\ &= -\frac{c_u}{c_v^2\sigma_v^2}\int_0^\infty g(u)u^{\alpha}\phi(-a)\,du + \left(\frac{\varepsilon}{c_v^2\sigma_v^2} - \frac{1}{c_u\sigma_u}\right)\times \\ &\qquad \int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du,\end{aligned}$$

$$\frac{\partial f(u)}{\partial\alpha} = \int_0^\infty \left(\frac{\partial g(u)}{\partial\alpha}u^{\alpha-1} + g(u)\ln u\cdot u^{\alpha-1}\right)\phi(-a)\,du,$$

$$\begin{aligned}\frac{\partial f(u)}{\partial(1/\sigma_u)} &= \int_0^\infty g(u)u^{\alpha-1}\left(-u+\frac{\varepsilon}{c_u}-\frac{c_v^2\sigma_v^2}{c_u^2\sigma_u}\right)\phi(-a)\,du \\ &= -\int_0^\infty g(u)u^{\alpha}\phi(-a)\,du+\left(\frac{\varepsilon}{c_u}-\frac{c_v^2\sigma_v^2}{c_u^2\sigma_u}\right)\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du,\end{aligned}$$

$$\begin{aligned}\frac{\partial f(u)}{\partial \sigma_v^2} &= \int_0^\infty g(u)u^{\alpha-1}\left(\frac{c_u^2}{2c_v^2\sigma_v^4}u^2-\frac{c_u\varepsilon}{c_v^2\sigma_v^4}u+\frac{\varepsilon^2}{2c_v^2\sigma_v^4}-\frac{c_v^2}{2c_u^2\sigma_u^2}\right)\phi(-a)\,du \\ &= \frac{c_u^2}{2c_v^2\sigma_v^4}\int_0^\infty g(u)u^{\alpha+1}\phi(-a)\,du-\frac{c_u\varepsilon}{c_v^2\sigma_v^4}\int_0^\infty g(u)u^{\alpha}\phi(-a)\,du \\ &\quad+\left(\frac{\varepsilon^2}{2c_v^2\sigma_v^4}-\frac{c_v^2}{2c_u^2\sigma_u^2}\right)\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du.\end{aligned}$$

Dividing these derivatives by $f(u)$ and letting $g(u)=1$ gives

$$\left(\frac{1}{f(u)}\cdot\frac{\partial f(u)}{\partial\boldsymbol{\beta}}\right)\Big|_{g(u)=1} = -\frac{c_u}{c_v^2\sigma_v^2}\mathbb{E}[U|E]+\frac{\varepsilon}{c_v^2\sigma_v^2}-\frac{1}{c_u\sigma_u},$$

$$\left(\frac{1}{f(u)}\cdot\frac{\partial f(u)}{\partial\alpha}\right)\Big|_{g(u)=1} = \mathbb{E}[\ln U|E],$$

$$\left(\frac{1}{f(u)}\cdot\frac{\partial f(u)}{\partial(1/\sigma_u)}\right)\Big|_{g(u)=1} = -\mathbb{E}[U|E]+\frac{\varepsilon}{c_u}-\frac{c_v^2\sigma_v^2}{c_u^2\sigma_u},$$

$$\left(\frac{1}{f(u)}\cdot\frac{\partial f(u)}{\partial\sigma_v^2}\right)\Big|_{g(u)=1} = \frac{c_u^2}{2c_v^2\sigma_v^4}\mathbb{E}[U^2|E]-\frac{c_u\varepsilon}{c_v^2\sigma_v^4}\mathbb{E}[U|E]+\frac{\varepsilon^2}{2c_v^2\sigma_v^4}-\frac{c_v^2}{2c_u^2\sigma_u^2},$$

where equation (C.11) in Section C.4 gives

$$\mathbb{E}[g(U)|E] = \frac{\displaystyle\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\displaystyle\int_0^\infty u^{\alpha-1}\phi(-a)\,du}.$$

The second-order derivatives of $\ln f(u)$ appear in the second-order derivatives of $\ln f_Y(y;\boldsymbol{\theta})$ under the normal-gamma specification in Section 2.5 of Chapter 2. Calculating these derivatives involves calculating the derivatives of $\mathbb{E}[g(U)|E]$,

which are given by

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\mathbb{E}[g(U)|E] &= -\left(\frac{c_u}{c_v^2\sigma_v^2}\right)\frac{\int_0^\infty g(u)u^\alpha\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du} \\
&\quad + \frac{c_u}{c_v^2\sigma_v^2}\left(\frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right)\left(\frac{\int_0^\infty u^{\alpha}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right) \\
&= -\frac{c_u}{c_v^2\sigma_v^2}\left\{\mathbb{E}[g(U)U|E] - \mathbb{E}[g(U)|E]\cdot\mathbb{E}[U|E]\right\} \\
&= -\frac{c_u}{c_v^2\sigma_v^2}Cov\left(g(U), U|E\right),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial f}{\partial \alpha}\mathbb{E}[g(U)|E] &= \frac{\int_0^\infty \frac{\partial g(u)}{\partial \alpha}u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du} + \frac{\int_0^\infty g(u)u^{\alpha-1}\ln u\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du} \\
&\quad - \left(\frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right)\left(\frac{\int_0^\infty u^{\alpha-1}\ln u\,\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right) \\
&= \mathbb{E}\left[\frac{\partial g(u)}{\partial \alpha}\middle| E\right] + \mathbb{E}[g(U)\ln U|E] - \mathbb{E}[g(U)|E]\cdot\mathbb{E}[\ln U|E] \\
&= \mathbb{E}\left[\frac{\partial g(u)}{\partial \alpha}\middle| E\right] + Cov\left(g(U), \ln U|E\right),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial(1/\sigma_u)}\mathbb{E}[g(U)|E] &= -\frac{\int_0^\infty g(u)u^\alpha\phi(-a)}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du} \\
&\quad + \left(\frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right)\left(\frac{\int_0^\infty u^{\alpha}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right) \\
&= -\mathbb{E}[g(U)U|E] + \mathbb{E}[g(U)|E]\cdot\mathbb{E}[U|E] \\
&= -Cov\left(g(U), U|E\right),
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \sigma_v^2}\mathbb{E}[g(U)|E] &= \frac{c_u^2}{2c_v^2\sigma_v^4}\frac{\int_0^\infty g(u)u^{\alpha+1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du} - \frac{c_u\varepsilon}{c_v^2\sigma_v^4}\frac{\int_0^\infty g(u)u^{\alpha}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du} \\
&\quad -\frac{c_u^2}{2c_v^2\sigma_v^4}\left(\frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right)\left(\frac{\int_0^\infty u^{\alpha+1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right) \\
&\quad +\frac{c_u\varepsilon}{c_v^2\sigma_v^4}\left(\frac{\int_0^\infty g(u)u^{\alpha-1}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right)\left(\frac{\int_0^\infty u^{\alpha}\phi(-a)\,du}{\int_0^\infty u^{\alpha-1}\phi(-a)\,du}\right) \\
&= \frac{c_u^2}{2c_v^2\sigma_v^4}\left\{\mathbb{E}[g(U)U^2|E] - \mathbb{E}[g(U)|E]\cdot\mathbb{E}[U^2|E]\right\} \\
&\quad -\frac{c_u\varepsilon}{c_v^2\sigma_v^4}\left\{\mathbb{E}[g(U)U|E] - \mathbb{E}[g(U)|E]\cdot\mathbb{E}[U|E]\right\} \\
&= \frac{c_u^2}{2c_v^2\sigma_v^4}Cov(g(U),U^2|E) - \frac{c_u\varepsilon}{c_v^2\sigma_v^4}Cov(g(U),U|E).
\end{aligned}
$$

Appendix B

Information Matrices for Stochastic Frontier Models

The information matrix for the log-linear Cobb-Douglas stochastic production frontier model can be obtained by letting $c_u = -1$, $c_v = 1$ and $f(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta}$ in the calculations for the information matrix for the general model (2.1). To reduce notational clutter, observation subscripts will be omitted.

B.1 Information Matrix for the Normal-Half Normal Model

The first-order partial derivatives of the log-likelihood function for a single observation are given by

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= \left\{ \frac{1}{\lambda\sigma_G} a + \frac{\lambda}{\sigma_G} h(a) \right\} \boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial \ln f_Y}{\partial \lambda} &= -\frac{1}{\lambda} a h(a),
\end{aligned}$$

$$\frac{\partial \ln f_Y}{\partial \sigma_G^2} = -\frac{1}{2\sigma_G^2} + \frac{1}{2}\left(\frac{1}{\lambda\sigma_G}\right)^2 a^2 + \frac{1}{2\sigma_G^2} ah(a).$$

The corresponding second-order partial derivatives are

$$\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \left\{-\frac{1}{\sigma_G^2} + \left(\frac{\lambda}{\sigma_G}\right)^2 [ah(a) - h(a)^2]\right\} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}),$$

$$\frac{\partial^2 \ln f_Y}{\partial \lambda^2} = \frac{1}{\lambda^2}[a^3 h(a) - a^2 h(a)^2],$$

$$\frac{\partial^2 \ln f_Y}{\partial (\sigma_G^2)^2} = \frac{1}{2\sigma_G^4} - \left(\frac{1}{\lambda\sigma_G^2}\right)^2 a^2 - \frac{1}{4\sigma_G^4}\left[3ah(a) - a^3 h(a) + a^2 h(a)^2\right],$$

$$\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \lambda} = \frac{1}{\sigma_G}\left\{h(a) - a^2 h(a) + ah(a)^2\right\} \boldsymbol{f}(\boldsymbol{x}),$$

$$\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \sigma_G^2} = \left\{-\frac{1}{\lambda\sigma_G^3} a - \frac{\lambda}{2\sigma_G^3}[h(a) - a^2 h(a) + ah(a)^2]\right\} \boldsymbol{f}(\boldsymbol{x}),$$

$$\frac{\partial^2 \ln f_Y}{\partial \lambda \partial \sigma_G^2} = \frac{1}{2\lambda\sigma_G^2}[ah(a) - a^3 h(a) + a^2 h(a)^2].$$

Using the first-order partial derivatives, the components of the per observation expected Fisher information matrix (2.12) are given by

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)^T\right] = \left\{\left(\frac{1}{\lambda\sigma_G}\right)^2 \mathbb{E}[a^2] + \frac{2}{\sigma_G^2}\mathbb{E}[ah(a)] + \left(\frac{\lambda}{\sigma_G}\right)^2 \mathbb{E}[h(a)^2]\right\} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)^2\right] = \frac{1}{\lambda^2}\mathbb{E}[a^2h(a)^2],$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)^2\right] = \left(\frac{1}{2\sigma_G^2}\right)^2 - \frac{1}{2}\left(\frac{1}{\lambda\sigma_G^2}\right)^2\mathbb{E}[a^2] - \frac{1}{2}\left(\frac{1}{\sigma_G^2}\right)^2\mathbb{E}[ah(a)] + \frac{1}{4}\left(\frac{1}{\lambda\sigma_G}\right)^4\mathbb{E}[a^4] + \frac{1}{2}\left(\frac{1}{\lambda\sigma_G^2}\right)^2\mathbb{E}[a^3h(a)] + \left(\frac{1}{2\sigma_G^2}\right)^2\mathbb{E}[a^2h(a)^2],$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\right] = \left\{-\frac{1}{\lambda^2\sigma_G}\mathbb{E}[a^2h(a)] - \frac{1}{\sigma_G}\mathbb{E}[ah(a)^2]\right\}\boldsymbol{f}(\boldsymbol{x}),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] = -\boldsymbol{f}(\boldsymbol{x})\left\{\frac{1}{2\lambda\sigma_G^3}\mathbb{E}[a] + \frac{1}{2}\left(-\frac{1}{\lambda\sigma_G}\right)^3\mathbb{E}[a^3] - \frac{1}{2\lambda\sigma_G^3}\mathbb{E}[a^2h(a)] + \frac{\lambda}{2\sigma_G^3}\mathbb{E}[h(a)] - \frac{1}{2\lambda\sigma_G^3}\mathbb{E}[a^2h(a)] - \frac{\lambda}{2\sigma_G^3}\mathbb{E}[ah(a)^2]\right\},$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] = \frac{1}{2\lambda\sigma_G^2}\mathbb{E}[ah(a)] - \frac{1}{2\lambda}\left(\frac{1}{\lambda\sigma_G}\right)^2\mathbb{E}[a^3h(a)] - \frac{1}{2\lambda\sigma_G^2}\mathbb{E}[a^2h(a)^2].$$

Using the second-order partial derivatives, the components of the per observation expected Fisher information matrix (2.13) are given by

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right] = -\left\{-\frac{1}{\sigma_G^2} + \left(\frac{\lambda}{\sigma_G}\right)^2\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right)\right\}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \lambda^2}\right] = -\frac{1}{\lambda^2}\left(\mathbb{E}[a^3h(a)] - \mathbb{E}[a^2h(a)^2]\right),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(\sigma_G^2)^2}\right] = -\frac{1}{2\sigma_G^4} + \left(\frac{1}{\lambda\sigma_G^2}\right)^2 \mathbb{E}[a^2] + \frac{1}{4\sigma_G^4}\left(3\mathbb{E}[ah(a)] - \mathbb{E}[a^3h(a)] + \mathbb{E}[a^2h(a)^2]\right),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\lambda}\right] = -\frac{1}{\sigma_G}\left\{\mathbb{E}[h(a)] - \mathbb{E}[a^2h(a)] + \mathbb{E}[ah(a)^2]\right\}\boldsymbol{f}(\boldsymbol{x}),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_G^2}\right] = -\left\{-\frac{1}{\lambda\sigma_G^3}\mathbb{E}[a] - \frac{\lambda}{2\sigma_G^3}\left(\mathbb{E}[h(a)] - \mathbb{E}[a^2h(a)] + \mathbb{E}[ah(a)^2]\right)\right\}\boldsymbol{f}(\boldsymbol{x}),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\lambda\partial\sigma_G^2}\right] = -\frac{1}{2\lambda\sigma_G^2}\left(\mathbb{E}[ah(a)] - \mathbb{E}[a^3h(a)] + \mathbb{E}[a^2h(a)^2]\right).$$

The following properties are useful in approximating the information matrix based on the approximations in Chapter 3

$$\mathbb{E}[a] = \frac{\lambda}{\sigma_G}\mathbb{E}[E],$$

$$Var(a) = \left(\frac{\lambda}{\sigma_G}\right)^2 Var(E),$$

$$\frac{\partial a}{\partial\boldsymbol{\beta}} = -\frac{\lambda}{\sigma_G}\boldsymbol{f}(\boldsymbol{x}).$$

B.2 Information Matrix for the Normal-Exponential Model

The first-order partial derivatives of the log-likelihood function for a single observation are given by

$$\frac{\partial \ln f_Y}{\partial\boldsymbol{\beta}} = -\left\{\frac{1}{\sigma_u} - \frac{1}{\sigma_v}h(a)\right\}\boldsymbol{f}(\boldsymbol{x}),$$

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial(1/\sigma_u)} &= \sigma_u + \sigma_v[a - h(a)], \\
\frac{\partial \ln f_Y}{\partial \sigma_v^2} &= \frac{1}{2\sigma_u^2} - \left(\frac{1}{\sigma_u\sigma_v} - \frac{1}{2\sigma_v^2}a\right) h(a),
\end{aligned}$$

The corresponding second-order partial derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \left\{\frac{1}{\sigma_v^2}[ah(a) - h(a)^2]\right\} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)^2} &= -\sigma_u^2 + \sigma_v^2[1 + ah(a) - h(a)^2], \\
\frac{\partial^2 \ln f_Y}{\partial(\sigma_v^2)^2} &= \frac{1}{\sigma_u^2\sigma_v^2}[ah(a) - h(a)^2] - \frac{1}{\sigma_u\sigma_v^3}[a^2h(a) - ah(a)^2 - h(a)] \\
&\quad + \frac{1}{4\sigma_v^4}[a^3h(a) - a^2h(a)^2 - 3ah(a)], \\
\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)} &= -\left\{1 + ah(a) - h(a)^2\right\} \boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_v^2} &= \left\{-\frac{1}{2\sigma_v^3}[h(a) - a^2h(a) + ah(a)^2] - \frac{1}{\sigma_u\sigma_v^2}[ah(a) - h(a)^2]\right\} \boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)\partial\sigma_v^2} &= \frac{1}{\sigma_u}[1 + ah(a) - h(a)^2] - \frac{1}{2\sigma_v}[h(a) + a^2h(a) - ah(a)^2].
\end{aligned}$$

Using the first-order partial derivatives, the components of the per observation expected Fisher information matrix (2.18) are given by

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial\boldsymbol{\beta}}\right)^T\right] = \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}) \times$$

$$\left\{\frac{1}{\sigma_u^2} - \frac{2}{\sigma_u\sigma_v}\mathbb{E}[h(a)] + \frac{1}{\sigma_v^2}\mathbb{E}[h(a)^2]\right\},$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial(1/\sigma_u)}\right)^2\right] = \sigma_u^2 + 2\sigma_u\sigma_v\left(\mathbb{E}[a] - \mathbb{E}[h(a)]\right) + \sigma_v^2\left(\mathbb{E}[a^2] - 2\mathbb{E}[ah(a)] + \mathbb{E}[h(a)^2]\right),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)^2\right] = \left(\frac{1}{2\sigma_u^2}\right)^2 - \frac{1}{\sigma_u^3\sigma_v}\mathbb{E}[h(a)] + \frac{1}{2\sigma_u^2\sigma_v^2}\mathbb{E}[ah(a)] + \left(\frac{1}{\sigma_u\sigma_v}\right)^2\mathbb{E}[h(a)^2] - \frac{1}{\sigma_u\sigma_v^3}\mathbb{E}[ah(a)^2] + \left(\frac{1}{2\sigma_v^2}\right)^2\mathbb{E}[a^2h(a)^2],$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial(1/\sigma_u)}\right)\right] = \boldsymbol{f}(\boldsymbol{x}) \times \left\{-\left(1 - \mathbb{E}[ah(a)] + \mathbb{E}[h(a)^2]\right) + \frac{\sigma_v}{\sigma_u}\left(\mathbb{E}[a] - \mathbb{E}[h(a)]\right) + \frac{\sigma_u}{\sigma_v}\mathbb{E}[h(a)]\right\},$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] = \left\{-\frac{1}{2\sigma_u^3} + \frac{1}{\sigma_u^2\sigma_v}\mathbb{E}[h(a)] - \frac{1}{2\sigma_u\sigma_v^2}\mathbb{E}[ah(a)] + \frac{1}{2\sigma_u^2\sigma_v}\mathbb{E}[h(a)] - \frac{1}{\sigma_u\sigma_v^2}\mathbb{E}[h(a)^2] + \frac{1}{2\sigma_v^3}\mathbb{E}[ah(a)^2]\right\}\boldsymbol{f}(\boldsymbol{x}),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial(1/\sigma_u)}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)\right] = \frac{1}{2\sigma_u} - \frac{1}{\sigma_v}\mathbb{E}[h(a)] + \frac{\sigma_u}{2\sigma_v^2}\mathbb{E}[ah(a)] + \frac{\sigma_v}{2\sigma_u^2}\left(\mathbb{E}[a] - \mathbb{E}[h(a)]\right) - \frac{1}{\sigma_u}\mathbb{E}[ah(a)] + \frac{1}{2\sigma_v}\mathbb{E}[a^2h(a)] + \frac{1}{\sigma_u}\mathbb{E}[h(a)^2] - \frac{1}{2\sigma_v}\mathbb{E}[ah(a)^2].$$

Using the second-order partial derivatives, the components of the per observation expected Fisher information matrix (2.19) are given by

$$
\begin{aligned}
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right] &= -\left\{\frac{1}{\sigma_v^2}\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right)\right\} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}), \\
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial (1/\sigma_u)^2}\right] &= \sigma_u^2 - \sigma_v^2\left(1 + \mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right), \\
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial (\sigma_v^2)^2}\right] &= -\frac{1}{\sigma_u^2\sigma_v^2}\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right) \\
&\quad + \frac{1}{\sigma_u\sigma_v^3}\left(\mathbb{E}[a^2h(a)] - \mathbb{E}[ah(a)^2] - \mathbb{E}[h(a)]\right) \\
&\quad - \frac{1}{4\sigma_v^4}\left(\mathbb{E}[a^3h(a)] - \mathbb{E}[a^2h(a)^2] - \mathbb{E}[3ah(a)]\right), \\
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial (1/\sigma_u)}\right] &= \left\{1 + \mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right\} \boldsymbol{f}(\boldsymbol{x}), \\
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \sigma_v^2}\right] &= \left\{\frac{1}{2\sigma_v^3}\left(\mathbb{E}[h(a)] - \mathbb{E}[a^2h(a)] + \mathbb{E}[ah(a)^2]\right)\right. \\
&\quad \left. \frac{1}{\sigma_u\sigma_v^2}\left(\mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right)\right\} \boldsymbol{f}(\boldsymbol{x}), \\
-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial (1/\sigma_u) \partial \sigma_v^2}\right] &= -\frac{1}{\sigma_u}\left(1 + \mathbb{E}[ah(a)] - \mathbb{E}[h(a)^2]\right) \\
&\quad + \frac{1}{2\sigma_v}\left(\mathbb{E}[h(a)] + \mathbb{E}[a^2h(a)] - \mathbb{E}[ah(a)^2]\right).
\end{aligned}
$$

The following properties are useful in approximating the information matrix based on the approximations in Chapter 3

$$
\mathbb{E}[a] = \frac{1}{\sigma_v}\mathbb{E}[E] + \frac{\sigma_v}{\sigma_u},
$$

$$Var(a) = \frac{1}{\sigma_v^2} Var(E),$$

$$\frac{\partial a}{\partial \boldsymbol{\beta}} = -\frac{1}{\sigma_v} \boldsymbol{f}(\boldsymbol{x}).$$

B.3 Information Matrix for the Normal-Truncated Normal Model

The first-order partial derivatives of the log-likelihood function for a single observation are given by

$$\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} = \left\{ \frac{\mu(\lambda^2+1)}{\lambda^2 \sigma_G^2} + \frac{1}{\lambda \sigma_G} a_1 + \frac{\lambda}{\sigma_G} h(a_1) \right\} \boldsymbol{f}(\boldsymbol{x}),$$

$$\frac{\partial \ln f_Y}{\partial \mu} = -\frac{\mu(\lambda^2+1)}{\lambda^2 \sigma_G^2} - \frac{1}{\lambda \sigma_G} a_1 + \frac{1}{\lambda \sigma_G} h(a_1) + \frac{a_2}{\mu} h(a_2),$$

$$\frac{\partial \ln f_Y}{\partial \lambda} = \left(-\frac{2\mu}{\lambda^2 \sigma_G} - \frac{1}{\lambda} a_1 \right) h(a_1) + \frac{\mu}{(\lambda^2+1)^{1/2} \lambda^2 \sigma_G} h(a_2),$$

$$\frac{\partial \ln f_Y}{\partial \sigma_G^2} = -\frac{1}{2\sigma_G^2} + \frac{1}{2} \left(\frac{\mu(\lambda^2+1)}{\lambda^2 \sigma_G^2} + \frac{1}{\lambda \sigma_G} a_1 \right)^2 + \frac{1}{2\sigma_G^2} a_1 h(a_1) - \frac{1}{2\sigma_G^2} a_2 h(a_2).$$

The corresponding second-order partial derivatives are

$$\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \left\{ -\frac{1}{\sigma_G^2} + \frac{\lambda^2}{\sigma_G^2} [a_1 h(a_1) - h(a_1)^2] \right\} \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{f}^T(\boldsymbol{x}),$$

$$\frac{\partial^2 \ln f_Y}{\partial \mu^2} = -\frac{1}{\sigma_G^2} + \frac{1}{\lambda^2 \sigma_G^2} [a_1 h(a_1) - h(a_1)^2] - \frac{a_2^2}{\mu^2} h(a_2)[a_2 - h(a_2)],$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \lambda^2} &= \frac{2\mu}{\lambda^3\sigma_G}h(a_1) + \frac{4\mu^2}{\lambda^4\sigma_G^2}[a_1h(a_1) - h(a_1)^2] + \frac{4\mu}{\lambda^3\sigma_G}[a_1^2h(a_1) - a_1h(a_1)^2] \\
&\quad + \frac{1}{\lambda^2}[a_1^3h(a_1) - a_1^2h(a_1)^2] - \frac{\mu(3\lambda^2+2)}{(\lambda^2+1)^{3/2}\lambda^3\sigma_G}h(a_2) \\
&\quad - \frac{\mu^2}{(\lambda^2+1)\lambda^4\sigma_G^2}h(a_2)[a_2 - h(a_2)], \\
\\
\frac{\partial^2 \ln f_Y}{\partial (\sigma_G^2)^2} &= -\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^6} - \frac{2\mu(\lambda^2+1)}{\lambda^3\sigma_G^5}a_1 - \frac{1}{\lambda^2\sigma_G^4}a_1^2 \\
&\quad + \frac{1}{4\sigma_G^4}\left(2 - 3a_1h(a_1) + a_1^3h(a_1) - a_1^2h(a_1)^2 + 3a_2h(a_2)\right. \\
&\quad \left. - a_2^2h(a_2)[a_2 - h(a_2)]\right), \\
\\
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta}\partial \mu} &= \frac{1}{\sigma_G^2}\{1 + a_1h(a_1) - h(a_1)^2\}\boldsymbol{f}(\boldsymbol{x}), \\
\\
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta}\partial \lambda} &= \frac{1}{\sigma_G}\left\{h(a_1) - \frac{2\mu}{\lambda\sigma_G}[a_1h(a_1) - h(a_1)^2] - a_1^2h(a_1) + a_1h(a_1)^2\right\}\boldsymbol{f}(\boldsymbol{x}), \\
\\
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta}\partial \sigma_G^2} &= \frac{1}{\sigma_G^3}\left\{-\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G} - \frac{1}{\lambda}a_1 - \frac{\lambda}{2}[h(a_1) - a_1^2h(a_1) + a_1h(a_1)^2]\right\}\boldsymbol{f}(\boldsymbol{x}), \\
\\
\frac{\partial^2 \ln f_Y}{\partial \mu\partial \lambda} &= -\frac{1}{\lambda^2\sigma_G}\left(h(a_1) + \frac{2\mu}{\lambda\sigma_G}[a_1h(a_1) - h(a_1)^2] + a_1^2h(a_1) - a_1h(a_1)^2\right) \\
&\quad + \frac{1}{(\lambda^2+1)^{1/2}\lambda^2\sigma_G}\left(h(a_2) - a_2h(a_2)[a_2 - h(a_2)]\right), \\
\\
\frac{\partial^2 \ln f_Y}{\partial \mu\partial \sigma_G^2} &= \frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^4} + \frac{1}{2\lambda\sigma_G^3}\left[2a_1 - h(a_1) + a_1^2h(a_1) - a_1h(a_1)^2\right] \\
&\quad - \frac{1}{2\mu\sigma_G^2}a_2h(a_2) + \frac{1}{2\mu\sigma_G^2}a_2^2h(a_2)[a_2 - h(a_2)], \\
\\
\frac{\partial^2 \ln f_Y}{\partial \lambda\partial \sigma_G^2} &= \frac{\mu}{\lambda^2\sigma_G^3}\left[h(a_1) - a_1^2h(a_1) + a_1h(a_1)^2\right] \\
&\quad + \frac{1}{2\lambda\sigma_G^2}\left[a_1h(a_1) - a_1^3h(a_1) + a_1^2h(a_1)^2\right]
\end{aligned}$$

$$-\frac{\mu}{2(\lambda^2+1)^{1/2}\lambda^2\sigma_G^3}\left(h(a_2)-a_2h(a_2)[a_2-h(a_2)]\right).$$

Using the first-order partial derivatives, the components of the per observation expected Fisher information matrix (2.26) are given by

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)^T\right] &= \boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x})\left\{\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^4}+\frac{2\mu(\lambda^2+1)}{\lambda^3\sigma_G^3}\mathbb{E}[a_1]\right.\\ &\left.+\frac{2\mu(\lambda^2+1)}{\lambda\sigma_G^3}\mathbb{E}[h(a_1)]+\frac{1}{\lambda^2\sigma_G^2}\mathbb{E}[a_1^2]+\frac{2}{\sigma_G^2}\mathbb{E}[a_1h(a_1)]+\frac{\lambda^2}{\sigma_G^2}\mathbb{E}[h(a_1)^2]\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \mu}\right)^2\right] &= \frac{a_2^2}{\mu^2}[a_2-h(a_2)]^2+\frac{2}{\mu\lambda\sigma_G}a_2[a_2-h(a_2)]\left(\mathbb{E}[a_1]-\mathbb{E}[h(a_1)]\right)\\ &+\frac{1}{\lambda^2\sigma_G^2}\left(\mathbb{E}[a_1^2]-2\mathbb{E}[a_1h(a_1)]+\mathbb{E}[h(a_1)^2]\right),\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)^2\right] &= \frac{4\mu^2}{\lambda^4\sigma_G^2}\mathbb{E}[h(a_1)^2]+\frac{4\mu}{\lambda^3\sigma_G}\mathbb{E}[a_1h(a_1)^2]\\ &-\frac{4\mu^2}{(\lambda^2+1)^{1/2}\lambda^4\sigma_G^2}h(a_2)\mathbb{E}[h(a_1)]+\frac{1}{\lambda^2}\mathbb{E}[a_1^2h(a_1)^2]\\ &-\frac{2\mu}{(\lambda^2+1)^{1/2}\lambda^3\sigma_G}h(a_2)\mathbb{E}[a_1h(a_1)]+\frac{\mu^2}{(\lambda^2+1)\lambda^4\sigma_G^2}h(a_2)^2,\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)^2\right] &= \frac{1}{4\sigma_G^4}\left[-\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^2}+a_2h(a_2)+1\right]^2\\ &-\frac{1}{\sigma_G^2}\left[-\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^2}+a_2h(a_2)+1\right]\times\\ &\left(\frac{\mu(\lambda^2+1)}{\lambda^3\sigma_G^3}\mathbb{E}[a_1]+\frac{1}{2\lambda^2\sigma_G^2}\mathbb{E}[a_1^2]+\frac{1}{2\sigma_G^2}\mathbb{E}[a_1h(a_1)]\right)\end{aligned}$$

$$+\frac{\mu^2(\lambda^2+1)^2}{\lambda^6\sigma_G^6}\mathbb{E}[a_1^2]+\frac{\mu(\lambda^2+1)}{\lambda^5\sigma_G^5}\mathbb{E}[a_1^3]+\frac{\mu(\lambda^2+1)}{\lambda^3\sigma_G^5}\mathbb{E}[a_1^2h(a_1)]$$
$$+\frac{1}{4\lambda^4\sigma_G^4}\mathbb{E}[a_1^4]+\frac{1}{2\lambda^2\sigma_G^4}\mathbb{E}[a_1^3h(a_1)]+\frac{1}{4\sigma_G^4}\mathbb{E}[a_1^2h(a_1)^2],$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \mu}\right)\right] &= \left\{\left[\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^2}-\frac{a_2}{\mu}h(a_2)\right]\times\right.\\ &\left(-\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^2}-\frac{1}{\lambda\sigma_G}\mathbb{E}[a_1]-\frac{\lambda}{\sigma_G}\mathbb{E}[h(a_1)]\right)\\ &-\frac{\mu(\lambda^2+1)}{\lambda^3\sigma_G^3}\left(\mathbb{E}[a_1]-\mathbb{E}[h(a_1)]\right)-\frac{1}{\lambda^2\sigma_G^2}\left(\mathbb{E}[a_1^2]-\mathbb{E}[a_1h(a_1)]\right)\\ &\left.-\frac{1}{\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)-h(a_1)^2]\right)\right\}\boldsymbol{f}(\boldsymbol{x}),\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\right] &= \left\{-\frac{2\mu^2(\lambda^2+1)}{\lambda^4\sigma_G^3}\mathbb{E}[h(a_1)]\right.\\ &-\frac{\mu(\lambda^2+1)}{\lambda^3\sigma_G^2}\mathbb{E}[a_1h(a_1)]+\frac{\mu^2(\lambda^2+1)^{1/2}}{\lambda^4\sigma_G^3}h(a_2)-\frac{2\mu}{\lambda^3\sigma_G^2}\mathbb{E}[a_1h(a_1)]\\ &-\frac{1}{\lambda^2\sigma_G}\mathbb{E}[a_1^2h(a_1)]+\frac{\mu}{(\lambda^2+1)^{1/2}\lambda^3\sigma_G^2}h(a_2)\mathbb{E}[a_1]-\frac{2\mu}{\lambda\sigma_G^2}\mathbb{E}[h(a_1)^2]\\ &\left.-\frac{1}{\sigma_G}\mathbb{E}[a_1h(a_1)^2]+\frac{\mu}{(\lambda^2+1)^{1/2}\lambda\sigma_G^2}h(a_2)\mathbb{E}[h(a_1)]\right\}\boldsymbol{f}(\boldsymbol{x}),\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] &= -\left\{-\frac{1}{2\sigma_G^2}\left[-\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^2}+a_2h(a_2)+1\right]\times\right.\\ &\left(-\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^2}-\frac{1}{\lambda\sigma_G}\mathbb{E}[a_1]-\frac{\lambda}{\sigma_G}\mathbb{E}[h(a_1)]\right)\\ &-\frac{\mu^2(\lambda^2+1)^2}{\lambda^5\sigma_G^5}\mathbb{E}[a_1]-\frac{3\mu(\lambda^2+1)}{2\lambda^4\sigma_G^4}\mathbb{E}[a_1^2]-\frac{3\mu(\lambda^2+1)}{2\lambda^2\sigma_G^4}\mathbb{E}[a_1h(a_1)]\\ &\left.-\frac{1}{2\lambda^3\sigma_G^3}\mathbb{E}[a_1^3]-\frac{1}{\lambda\sigma_G^3}\mathbb{E}[a_1^2h(a_1)]-\frac{\lambda}{2\sigma_G^3}\mathbb{E}[a_1h(a_1)^2]\right\}\boldsymbol{f}(\boldsymbol{x}),\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \mu}\right)\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\right] &= \left[\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^2} - \frac{a_2}{\mu}h(a_2)\right] \times \\
&\left(\frac{2\mu}{\lambda^2\sigma_G}\mathbb{E}[h(a_1)] + \frac{1}{\lambda}\mathbb{E}[a_1h(a_1)] - \frac{\mu}{(\lambda^2+1)^{1/2}\lambda^2\sigma_G}h(a_2)\right) \\
&+\frac{2\mu}{\lambda^3\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)] - \mathbb{E}[h(a_1)^2]\right) + \frac{1}{\lambda^2\sigma_G}\left(\mathbb{E}[a_1^2h(a_1)] - \mathbb{E}[a_1h(a_1)^2]\right) \\
&-\frac{\mu}{(\lambda^2+1)^{1/2}\lambda^3\sigma_G^2}h(a_2)\left(\mathbb{E}[a_1] - \mathbb{E}[h(a_1)]\right),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \mu}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] &= \left[\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^2} - \frac{a_2}{\mu}h(a_2)\right] \times \\
&\left(\frac{1}{2\sigma_G^2}\left[-\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^2} + a_2h(a_2) + 1\right] - \frac{\mu(\lambda^2+1)}{\lambda^3\sigma_G^3}\mathbb{E}[a_1]\right. \\
&\left.-\frac{1}{2\lambda^2\sigma_G^2}\mathbb{E}[a_1^2] - \frac{1}{2\sigma_G^2}\mathbb{E}[a_1h(a_1)]\right) \\
&+\frac{1}{2\lambda\sigma_G^3}\left[-\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^2} + a_2h(a_2) + 1\right]\left(\mathbb{E}[a_1] - \mathbb{E}[h(a_1)]\right) \\
&-\frac{\mu(\lambda^2+1)}{\lambda^4\sigma_G^4}\left(\mathbb{E}[a_1^2] - \mathbb{E}[a_1h(a_1)]\right) - \frac{1}{2\lambda^3\sigma_G^3}\left(\mathbb{E}[a_1^3] - \mathbb{E}[a_1^2h(a_1)]\right) \\
&-\frac{1}{2\lambda\sigma_G^3}\left(\mathbb{E}[a_1^2h(a_1)] - \mathbb{E}[a_1h(a_1)^2]\right),
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \lambda}\right)\left(\frac{\partial \ln f_Y}{\partial \sigma_G^2}\right)\right] &= \frac{1}{2\sigma_G^2}\left[-\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^2} + a_2h(a_2) + 1\right] \times \\
&\left(\frac{2\mu}{\lambda^2\sigma_G}\mathbb{E}[h(a_1)] + \frac{1}{\lambda}\mathbb{E}[a_1h(a_1)] - \frac{\mu}{(\lambda^2+1)^{1/2}\lambda^2\sigma_G}h(a_2)\right) \\
&-\frac{2\mu^2(\lambda^2+1)}{\lambda^5\sigma_G^4}\mathbb{E}[a_1h(a_1)] - \frac{\mu}{\lambda^4\sigma_G^3}\mathbb{E}[a_1^2h(a_1)] - \frac{\mu}{\lambda^2\sigma_G^3}\mathbb{E}[a_1h(a_1)^2] \\
&-\frac{\mu(\lambda^2+1)}{\lambda^4\sigma_G^3}\mathbb{E}[a_1^2h(a_1)] - \frac{1}{2\lambda^3\sigma_G^2}\mathbb{E}[a_1^3h(a_1)] - \frac{1}{2\lambda\sigma_G^2}\mathbb{E}[a_1^2h(a_1)^2] \\
&+\frac{\mu}{(\lambda^2+1)^{1/2}\lambda^2\sigma_G}h(a_2) \times
\end{aligned}$$

$$\left(\frac{\mu(\lambda^2+1)}{\lambda^3\sigma_G^3}\mathbb{E}[a_1]+\frac{1}{2\lambda^2\sigma_G^2}\mathbb{E}[a_1^2]+\frac{1}{2\sigma_G^2}\mathbb{E}[a_1h(a_1)]\right).$$

Using the second-order partial derivatives, the components of the per observation expected Fisher information matrix (2.27) are given by

$$-\mathbb{E}\left[\frac{\partial^2\ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = -\left\{-\frac{1}{\sigma_G^2}+\frac{\lambda^2}{\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right)\right\}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}),$$

$$-\mathbb{E}\left[\frac{\partial^2\ln f_Y}{\partial\mu^2}\right] = \frac{1}{\sigma_G^2}-\frac{1}{\lambda^2\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right)+\frac{a_2^2}{\mu^2}h(a_2)[a_2-h(a_2)],$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2\ln f_Y}{\partial\lambda^2}\right] = &-\frac{2\mu}{\lambda^3\sigma_G}\mathbb{E}[h(a_1)]-\frac{4\mu^2}{\lambda^4\sigma_G^2}\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right)\\ &-\frac{4\mu}{\lambda^3\sigma_G}\left(\mathbb{E}[a_1^2h(a_1)]-\mathbb{E}[a_1h(a_1)^2]\right)\\ &-\frac{1}{\lambda^2}\left(\mathbb{E}[a_1^3h(a_1)]-\mathbb{E}[a_1^2h(a_1)^2]\right)+\frac{\mu(3\lambda^2+2)}{(\lambda^2+1)^{3/2}\lambda^3\sigma_G}h(a_2)\\ &+\frac{\mu^2}{(\lambda^2+1)\lambda^4\sigma_G^2}h(a_2)[a_2-h(a_2)],\end{aligned}$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2\ln f_Y}{\partial(\sigma_G^2)^2}\right] = &\frac{\mu^2(\lambda^2+1)^2}{\lambda^4\sigma_G^6}+\frac{2\mu(\lambda^2+1)}{\lambda^3\sigma_G^5}\mathbb{E}[a_1]+\frac{1}{\lambda^2\sigma_G^4}\mathbb{E}[a_1^2]\\ &-\frac{1}{4\sigma_G^4}\left(2-3\mathbb{E}[a_1h(a_1)]+\mathbb{E}[a_1^3h(a_1)]-\mathbb{E}[a_1^2h(a_1)^2]\right.\\ &\left.+3a_2h(a_2)-a_2^2h(a_2)[a_2-h(a_2)]\right),\end{aligned}$$

$$-\mathbb{E}\left[\frac{\partial^2\ln f_Y}{\partial\boldsymbol{\beta}\partial\mu}\right] = -\frac{1}{\sigma_G^2}\{1+\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\}\boldsymbol{f}(\boldsymbol{x}),$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2\ln f_Y}{\partial\boldsymbol{\beta}\partial\lambda}\right] = &\boldsymbol{f}(\boldsymbol{x})\times\\ &-\frac{1}{\sigma_G}\left\{\mathbb{E}[h(a_1)]-\frac{2\mu}{\lambda\sigma_G}\left(\mathbb{E}[a_1h(a_1)]-\mathbb{E}[h(a_1)^2]\right)-\mathbb{E}[a_1^2h(a_1)]+\mathbb{E}[a_1h(a_1)^2]\right\},\end{aligned}$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \sigma_G^2}\right] = -\frac{1}{\sigma_G^3}\boldsymbol{f}(\boldsymbol{x}) \times \left\{-\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G} - \frac{1}{\lambda}\mathbb{E}[a_1] - \frac{\lambda}{2}\left(\mathbb{E}[h(a_1)] - \mathbb{E}[a_1^2 h(a_1)] + \mathbb{E}[a_1 h(a_1)^2]\right)\right\},$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \mu \partial \lambda}\right] = \frac{1}{\lambda^2\sigma_G}\left(\mathbb{E}[h(a_1)] + \frac{2\mu}{\lambda\sigma_G}\left(\mathbb{E}[a_1 h(a_1)] - \mathbb{E}[h(a_1)^2]\right) + \mathbb{E}[a_1^2 h(a_1)] - \mathbb{E}[a_1 h(a_1)^2]\right) - \frac{1}{(\lambda^2+1)^{1/2}\lambda^2\sigma_G}\left(h(a_2) - a_2 h(a_2)[a_2 - h(a_2)]\right),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \mu \partial \sigma_G^2}\right] = -\frac{\mu(\lambda^2+1)}{\lambda^2\sigma_G^4} - \frac{1}{2\lambda\sigma_G^3}\left(2\mathbb{E}[a_1] - \mathbb{E}[h(a_1)] + \mathbb{E}[a_1^2 h(a_1)] - \mathbb{E}[a_1 h(a_1)^2]\right) + \frac{1}{2\mu\sigma_G^2}a_2 h(a_2) - \frac{1}{2\mu\sigma_G^2}a_2^2 h(a_2)[a_2 - h(a_2)],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial \lambda \partial \sigma_G^2}\right] = -\frac{\mu}{\lambda^2\sigma_G^3}\left(\mathbb{E}[h(a_1)] - \mathbb{E}[a_1^2 h(a_1)] + \mathbb{E}[a_1 h(a_1)^2]\right) - \frac{1}{2\lambda\sigma_G^2}\left(\mathbb{E}[a_1 h(a_1)] - \mathbb{E}[a_1^3 h(a_1)] + \mathbb{E}[a_1^2 h(a_1)^2]\right) + \frac{\mu}{2(\lambda^2+1)^{1/2}\lambda^2\sigma_G^3}\left(h(a_2) - a_2 h(a_2)[a_2 - h(a_2)]\right).$$

The following properties are useful in approximating the information matrix based on the approximations in Chapter 3

$$\mathbb{E}[a_1] = -\frac{\mu}{\lambda\sigma_G} + \frac{\lambda}{\sigma_G}\mathbb{E}[E],$$

$$Var(a_1) = \frac{\lambda^2}{\sigma_G^2}Var(E),$$

$$\frac{\partial a_1}{\partial \boldsymbol{\beta}} = -\frac{\lambda}{\sigma_G}\boldsymbol{f}(\boldsymbol{x}).$$

When $\mu = 0$ the nonnegative truncated normal distribution collapses to the

nonnegative half normal distribution and substituting $\mu = 0$ into the normal-truncated normal equations above gives the corresponding normal-half normal equations in Appendix B.1.

B.4 Information Matrix for the Normal-Gamma Model

The first-order partial derivatives of the log-likelihood function for a single observation are given by

$$\begin{aligned}
\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma_v^2}\left\{\varepsilon + \mathbb{E}[U|E]\right\}\boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial \ln f_Y}{\partial \alpha} &= -\psi(\alpha) + \ln\left(\frac{1}{\sigma_u}\right) + \mathbb{E}[\ln U|E], \\
\frac{\partial \ln f_Y}{\partial (1/\sigma_u)} &= \alpha\sigma_u - \mathbb{E}[U|E], \\
\frac{\partial \ln f_Y}{\partial \sigma_v^2} &= -\frac{1}{2\sigma_v^2} + \frac{1}{2\sigma_v^4}\left(\varepsilon^2 + \mathbb{E}[U^2|E] + 2\varepsilon\mathbb{E}[U|E]\right),
\end{aligned}$$

where $\psi(\alpha)$ is the digamma function and $\mathbb{E}[g(U)|E] = \dfrac{\mathbb{E}[g(Q)Q^{\alpha-1}]}{\mathbb{E}[Q^{\alpha-1}]}$ for any given function g of U. The corresponding second-order partial derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= \frac{1}{\sigma_v^2}\left\{-1 + \frac{1}{\sigma_v^2}Var(U|E)\right\}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial \alpha^2} &= -\psi_1(\alpha) + Var(\ln U|E),
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \ln f_Y}{\partial (1/\sigma_u)^2} &= -\alpha\sigma_u^2 + Var(U|E), \\
\frac{\partial^2 \ln f_Y}{\partial (\sigma_v^2)^2} &= \frac{1}{2\sigma_v^4} - \frac{1}{\sigma_v^6}\left(\varepsilon^2 + \mathbb{E}[U^2|E] + 2\varepsilon\mathbb{E}[U|E]\right) \\
&\quad + \frac{1}{\sigma_v^8}\left[\frac{1}{4}Var(U^2|E) + \varepsilon Cov(U, U^2|E) + \varepsilon^2 Var(U|E)\right], \\
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \alpha} &= \frac{1}{\sigma_v^2} Cov(U, \ln U|E)\boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial (1/\sigma_u)} &= -\frac{1}{\sigma_v^2} Var(U|E)\boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial \boldsymbol{\beta} \partial \sigma_v^2} &= -\frac{1}{\sigma_v^4}\left\{\varepsilon + \mathbb{E}[U|E] - \frac{1}{2\sigma_v^2} Cov(U, U^2|E) - \frac{1}{\sigma_v^2}\varepsilon Var(U|E)\right\}\boldsymbol{f}(\boldsymbol{x}), \\
\frac{\partial^2 \ln f_Y}{\partial \alpha \partial (1/\sigma_u)} &= \sigma_u - Cov(U, \ln U|E), \\
\frac{\partial^2 \ln f_Y}{\partial \alpha \partial \sigma_v^2} &= \frac{1}{2\sigma_v^4} Cov(U^2, \ln U|E) + \frac{1}{\sigma_v^4}\varepsilon Cov(U, \ln U|E), \\
\frac{\partial^2 \ln f_Y}{\partial (1/\sigma_u) \partial \sigma_v^2} &= -\frac{1}{2\sigma_v^4} Cov(U, U^2|E) - \frac{1}{\sigma_v^4}\varepsilon Var(U|E),
\end{aligned}$$

where $\psi_1(\alpha)$ is the trigamma function and

$$\begin{aligned}
Var(g(U)|E) &= \mathbb{E}[g(U)^2|E] - \mathbb{E}[g(U)|E]^2, \\
Cov(f(U), g(U)|E) &= \mathbb{E}[f(U) \cdot g(U)|E] - \mathbb{E}[f(U)|E] \cdot \mathbb{E}[g(U)|E],
\end{aligned}$$

for given functions f and g of U. Using the first-order partial derivatives, the components of the per observation expected Fisher information matrix (2.33) are given by

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)^T\right] = \frac{1}{\sigma_v^4}\left\{\mathbb{E}[E^2] + 2\mathbb{E}(E \cdot \mathbb{E}[U|E]) + \mathbb{E}(\mathbb{E}[U|E]^2)\right\}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \alpha}\right)^2\right] = \left[-\psi(\alpha) - \ln \sigma_u\right]^2 + 2\left[-\psi(\alpha) - \ln \sigma_u\right]\mathbb{E}\left(\mathbb{E}[\ln U|E]\right) + \mathbb{E}\left(\mathbb{E}[\ln U|E]^2\right),$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)^2\right] = \alpha^2\sigma_u^2 - 2\alpha\sigma_u\mathbb{E}\left(\mathbb{E}[U|E]\right) + \mathbb{E}\left(\mathbb{E}[U|E]^2\right),$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \sigma_v^2}\right)^2\right] = {} & \frac{1}{4\sigma_v^4} - \frac{1}{2\sigma_v^6}\left\{\mathbb{E}[E^2]\mathbb{E}\left(\mathbb{E}[U^2|E]\right) + 2\mathbb{E}\left(E \cdot \mathbb{E}[U|E]\right)\right\} \\ & + \frac{1}{4\sigma_v^8}\left\{\mathbb{E}[E^4] + 2\mathbb{E}\left(E^2 \cdot \mathbb{E}[U^2|E]\right) + 4\mathbb{E}\left(E^3 \cdot \mathbb{E}[U|E]\right)\right. \\ & \left. + \mathbb{E}\left(\mathbb{E}[U^2|E]^2\right) + 2\mathbb{E}\left(E \cdot \mathbb{E}[U|E] \cdot \mathbb{E}[U^2|E]\right) + 4\mathbb{E}\left(E^2 \cdot \mathbb{E}[U|E]^2\right)\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial \alpha}\right)\right] = {} & \frac{1}{\sigma_v^2}\left\{\left[-\psi(\alpha) - \ln \sigma_u\right]\left(\mathbb{E}[E] + \mathbb{E}[U|E]\right)\right. \\ & \left. + \mathbb{E}\left(E \cdot \mathbb{E}[\ln U|E]\right) - \mathbb{E}\left(\mathbb{E}[U|E] \cdot \mathbb{E}[\ln U|E]\right)\right\}\boldsymbol{f}(\boldsymbol{x}),\end{aligned}$$

$$\mathbb{E}\left[\left(\frac{\partial \ln f_Y}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_Y}{\partial (1/\sigma_u)}\right)\right] =$$

$$\frac{1}{\sigma_v^2}\left\{\alpha\sigma_u\mathbb{E}[E]+\alpha\sigma_u\mathbb{E}[U|E]-\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)-\mathbb{E}\left(\mathbb{E}[U|E]^2\right)\right\}\boldsymbol{f}(\boldsymbol{x}),$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial\ln f_Y}{\partial\sigma_v^2}\right)\right] &= -\frac{1}{2\sigma_v^4}\left\{\mathbb{E}[E]+\mathbb{E}\left(\mathbb{E}[U|E]\right)\right\}\\ &+\frac{1}{2\sigma_v^6}\left\{\mathbb{E}[E^3]+\mathbb{E}\left(E\cdot\mathbb{E}[U^2|E]\right)+2\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]\right)\right\}\\ &+\frac{1}{2\sigma_v^6}\left\{\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]\right)+\mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[U^2|E]\right)+2\mathbb{E}\left(E\cdot\mathbb{E}[U|E]^2\right)\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\alpha}\right)\left(\frac{\partial\ln f_Y}{\partial(1/\sigma_u)}\right)\right] &= \left[-\psi(\alpha)-\ln\sigma_u\right]\left\{\alpha\sigma_u-\mathbb{E}\left(\mathbb{E}[U|E]\right)\right\}\\ &+\alpha\sigma_u\mathbb{E}\left(\mathbb{E}[\ln U|E]\right)-\mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[\ln U|E]\right),\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial\alpha}\right)\left(\frac{\partial\ln f_Y}{\partial\sigma_v^2}\right)\right] &= -\frac{1}{2\sigma_v^2}\left\{-\psi(\alpha)-\ln\sigma_u+\mathbb{E}\left(\mathbb{E}[\ln U|E]\right)\right\}\\ &+\frac{1}{2\sigma_v^4}\left[-\psi(\alpha)-\ln\sigma_u\right]\left\{\mathbb{E}[E^2]+\mathbb{E}\left(\mathbb{E}[U^2|E]\right)+2\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\}\\ &+\frac{1}{2\sigma_v^4}\left\{\mathbb{E}\left(E^2\cdot\mathbb{E}[\ln U|E]\right)+\mathbb{E}\left(\mathbb{E}[U^2|E]\cdot\mathbb{E}[\ln U|E]\right)\right.\\ &\left.+2\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\cdot\mathbb{E}[\ln U|E]\right)\right\},\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\partial\ln f_Y}{\partial(1/\sigma_u)}\right)\left(\frac{\partial\ln f_Y}{\partial\sigma_v^2}\right)\right] &= -\frac{1}{2\sigma_v^2}\left\{\alpha\sigma_u-\mathbb{E}\left(\mathbb{E}[U|E]\right)\right\}\\ &+\frac{\alpha\sigma_u}{2\sigma_v^4}\left\{\mathbb{E}[E^2]+\mathbb{E}\left(\mathbb{E}[U^2|E]\right)+2\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\}\\ &-\frac{1}{2\sigma_v^4}\left\{\mathbb{E}\left(E^2\cdot\mathbb{E}[U|E]\right)+\mathbb{E}\left(\mathbb{E}[U|E]\cdot\mathbb{E}[U^2|E]\right)+2\mathbb{E}\left(E\cdot\mathbb{E}[U|E]^2\right)\right\}.\end{aligned}$$

Using the second-order partial derivatives, the components of the per observation expected Fisher information matrix (2.32) are given by

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = -\frac{1}{\sigma_v^2}\left\{-1+\frac{1}{\sigma_v^2}\mathbb{E}\left[Var(U|E)\right]\right\}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{f}^T(\boldsymbol{x}),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\alpha^2}\right] = \psi_1(\alpha)-\mathbb{E}\left[Var(\ln U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)^2}\right] = \alpha\sigma_u^2-\mathbb{E}\left[Var(U|E)\right],$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(\sigma_v^2)^2}\right] &= -\frac{1}{2\sigma_v^4}+\frac{1}{\sigma_v^6}\left\{\mathbb{E}[E^2]+\mathbb{E}\left(\mathbb{E}[U^2|E]\right)+2\mathbb{E}\left(E\cdot\mathbb{E}[U|E]\right)\right\}\\ &\quad -\frac{1}{\sigma_v^8}\left\{\frac{1}{4}\mathbb{E}\left[Var(U^2|E)\right]+\mathbb{E}\left[E\cdot Cov(U,U^2|E)\right]+\mathbb{E}\left[E^2\cdot Var(U|E)\right]\right\},\end{aligned}$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\alpha}\right] = -\frac{1}{\sigma_v^2}\mathbb{E}\left[Cov(U,\ln U|E)\right]\boldsymbol{f}(\boldsymbol{x}),$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial(1/\sigma_u)}\right] = \frac{1}{\sigma_v^2}\mathbb{E}\left[Var(U|E)\right]\boldsymbol{f}(\boldsymbol{x}),$$

$$\begin{aligned}-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\boldsymbol{\beta}\partial\sigma_v^2}\right] &= \frac{1}{\sigma_v^4}\boldsymbol{f}(\boldsymbol{x})\times\\ &\quad\left\{\mathbb{E}[E]+\mathbb{E}\left(\mathbb{E}[U|E]\right)+\frac{1}{2\sigma_v^2}\mathbb{E}\left[Cov(U,U^2|E)\right]-\frac{1}{\sigma_v^2}\mathbb{E}\left[E\cdot Var(U|E)\right]\right\},\end{aligned}$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\alpha\partial(1/\sigma_u)}\right] = -\sigma_u+\mathbb{E}\left[Cov(U,\ln U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial\alpha\partial\sigma_v^2}\right] = -\frac{1}{2\sigma_v^4}\mathbb{E}\left[Cov(U^2,\ln U|E)\right]-\frac{1}{\sigma_v^4}\mathbb{E}\left[E\cdot Cov(U,\ln U|E)\right],$$

$$-\mathbb{E}\left[\frac{\partial^2 \ln f_Y}{\partial(1/\sigma_u)\partial\sigma_v^2}\right] = \frac{1}{2\sigma_v^4}\mathbb{E}\left[Cov(U,U^2|E)\right]+\frac{1}{\sigma_v^4}\mathbb{E}\left[E\cdot Var(U|E)\right].$$

Along with the expected value and variance of E, the following property is useful in approximating the information matrix based on the approximations in Chapter 3

$$\frac{\partial \varepsilon}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ y - \boldsymbol{f}^T(\boldsymbol{x})\boldsymbol{\beta} \right\} = -\boldsymbol{f}(\boldsymbol{x}).$$

When $\alpha = 1$ the gamma distribution collapses to the exponential distribution and substituting $\alpha = 1$ into the normal-gamma equations above gives the corresponding normal-exponential equations in Section 4.3.2 (after some further algebraic manipulation).

Appendix C

Ancillary Equations

C.1 Method for Obtaining the Joint Density $f_{U,E}(u, \varepsilon)$

The joint probability density function of random variables U and E can be derived by considering a change (or transformation) of variables. Let U and V be independent random variables with respective probability density functions $f_U(u)$ and $f_V(v)$. Let

$$\begin{aligned} u &= g_1(u,v) &&= u, \\ \varepsilon &= g_2(u,v) &&= c_u u + c_v v, \end{aligned}$$

define a one-to-one continuously differentiable transformation with inverse

$$\begin{aligned} u &= h_1(u,\varepsilon) &&= u, \\ v &= h_2(u,\varepsilon) &&= \frac{\varepsilon - c_u u}{c_v}, \end{aligned}$$

where $\{c_u, c_v\} \in \mathbb{R}$. The determinant of order 2,

$$J(u,\varepsilon) = \det\left(\frac{\partial(u,v)}{\partial(u,\varepsilon)}\right) = \begin{vmatrix} \frac{\partial u}{\partial u} & \frac{\partial u}{\partial \varepsilon} \\ \frac{\partial v}{\partial u} & \frac{\partial v}{\partial \varepsilon} \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -\frac{c_u}{c_v} & \frac{1}{c_v} \end{vmatrix} = \frac{1}{c_v},$$

is the Jacobian of the transformation. The joint density of U and E is given by

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= f_U(h_1(u,\varepsilon)) \times f_V(h_2(u,\varepsilon)) \times |J(u,\varepsilon)| \\ &= f_U(u) \times f_V\left(\frac{\varepsilon - c_u u}{c_v}\right) \times \left|\frac{1}{c_v}\right| \\ &= \frac{1}{|c_v|} f_{U,V}\left(u, \frac{\varepsilon - c_u u}{c_v}\right). \end{aligned} \tag{C.1}$$

C.2 Method for Obtaining the Marginal Density $f_E(\varepsilon)$

For random variables U and E with joint density of the form

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= u^{\alpha-1} K \exp\left\{-\frac{1}{2}[Au^2 - 2Bu + C]\right\}, \\ u \geq 0,\ -\infty < \varepsilon < \infty,\ \alpha > 0, \end{aligned} \tag{C.2}$$

the marginal density of E can be obtained by integrating u out of $f_{U,E}(u,\varepsilon)$. The coefficients K, A, B and C are functions of various parameters which are not discussed here. The integration of $f_{U,E}(u,\varepsilon)$ with respect to u can be easily calculated by first completing the square of equation (C.2) as follows

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= u^{\alpha-1} K \exp\left\{-\frac{1}{2}[Au^2 - 2Bu + C]\right\} \\ &= u^{\alpha-1} K \exp\left\{-\frac{1}{2}\left[A\left(u^2 - 2\frac{B}{A}u\right) + C\right]\right\} \\ &= u^{\alpha-1} K \exp\left\{-\frac{1}{2}\left[A\left(u - \frac{B}{A}\right)^2 + C - \frac{B^2}{A}\right]\right\} \\ &= u^{\alpha-1} K \exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \exp\left\{-\frac{A}{2}\left(u - \frac{B}{A}\right)^2\right\} \\ &= u^{\alpha-1} K \exp\left\{-\frac{1}{2}\left(C - \frac{B^2}{A}\right)\right\} \exp\left\{-\frac{1}{2}\left(\frac{u - B/A}{1/\sqrt{A}}\right)^2\right\}. \end{aligned} \tag{C.3}$$

If the coefficients K, A, B and C are not functions of u then any terms involving these coefficients (and not u) can be taken outside of the integration. The marginal density of E is then given by

$$\begin{aligned} f_E(\varepsilon) &= \int_0^\infty f_{U,E}(u,\varepsilon)\,du \\ &= K\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\int_0^\infty u^{\alpha-1}\exp\left\{-\frac{1}{2}\left(\frac{u-B/A}{1/\sqrt{A}}\right)^2\right\}du. \end{aligned} \tag{C.4}$$

Multiplying and dividing the above equation by both $\sqrt{\frac{2\pi}{A}}$ and $\Phi\left(\frac{B/A}{1/\sqrt{A}}\right)$ gives

$$f_E(\varepsilon) = K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B/A}{1/\sqrt{A}}\right)\times \frac{\int_0^\infty u^{\alpha-1}\sqrt{\frac{A}{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{u-B/A}{1/\sqrt{A}}\right)^2\right\}du}{\Phi\left(\frac{B/A}{1/\sqrt{A}}\right)},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Let random variable Q have a normal distribution, with mean B/A and variance $1/A$, which is truncated from below at zero, i.e. $Q \sim N^+\left(\frac{B}{A},\frac{1}{A}\right)$, then

$$f_Q\left(q;\frac{B}{A},\frac{1}{\sqrt{A}}\right) = \frac{\sqrt{\frac{A}{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{q-B/A}{1/\sqrt{A}}\right)^2\right\}}{\Phi\left(\frac{B/A}{1/\sqrt{A}}\right)}, \qquad q\geq 0,$$

is the probability density function of Q and

$$\mathbb{E}[Q^{\alpha-1}] = \int_0^\infty q^{\alpha-1} f_Q\left(q;\frac{B}{A},\frac{1}{\sqrt{A}}\right)dq$$

is a fractional moment of the nonnegative truncated normal distribution of Q. Appendix C.5 provides further details on truncated normal distributions.

Thus the marginal density of E can be expressed as

$$\begin{aligned} f_E(\varepsilon) &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B/A}{1/\sqrt{A}}\right)\int_0^\infty q^{\alpha-1}f_Q\left(q;\frac{B}{A},\frac{1}{\sqrt{A}}\right)dq \\ &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B/A}{1/\sqrt{A}}\right)\mathbb{E}[Q^{\alpha-1}] \\ &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B}{\sqrt{A}}\right)\mathbb{E}[Q^{\alpha-1}]. \end{aligned} \tag{C.5}$$

When $\alpha = 1$, the respective joint and marginal densities are

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= K\exp\left\{-\frac{1}{2}[Au^2-2Bu+C]\right\} \\ &= K\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\exp\left\{-\frac{1}{2}\left(\frac{u-B/A}{1/\sqrt{A}}\right)^2\right\}, \end{aligned}$$

$$f_E(\varepsilon) = K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B}{\sqrt{A}}\right). \tag{C.6}$$

C.3 Expected Value and Variance of E

Suppose that random variable V has a normal distribution with $\mathbb{E}[V] = 0$ and $Var(V) = \sigma_v^2$, and that random variable U has an unspecified distribution. If U and V are independent and $E = c_u U + c_v V$ where $\{c_u, c_v\} \in \mathbb{R}$, then the expected value and variance of E are

$$\begin{aligned} \mathbb{E}[E] &= \mathbb{E}[c_u U + c_v V] \\ &= c_u\mathbb{E}[U] + c_v\mathbb{E}[V] \\ &= c_u\mathbb{E}[U], \end{aligned} \tag{C.7}$$

$$\begin{aligned} Var(E) &= Var(c_u U + c_v V) \\ &= c_u^2 Var(U) + c_v^2 Var(V) \\ &= c_u^2 Var(U) + c_v^2\sigma_v^2. \end{aligned} \tag{C.8}$$

C.4 Conditional Density $f_{U|E}(u|\varepsilon)$

If the joint density of U and E is of the form given in equation (C.3)

$$\begin{aligned} f_{U,E}(u,\varepsilon) &= u^{\alpha-1}K\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\exp\left\{-\frac{1}{2}\left(\frac{u-B/A}{1/\sqrt{A}}\right)^2\right\} \\ &= u^{\alpha-1}K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right), \end{aligned}$$

$$u \geq 0,\ -\infty < \varepsilon < \infty,\ \alpha > 0,$$

and the marginal density of E is of the form given in equations (C.4) and (C.5)

$$\begin{aligned} f_E(\varepsilon) &= K\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\int_0^\infty u^{\alpha-1}\exp\left\{-\frac{1}{2}\left(\frac{u-B/A}{1/\sqrt{A}}\right)^2\right\}du \\ &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du \\ &= K\sqrt{\frac{2\pi}{A}}\exp\left\{-\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right\}\Phi\left(\frac{B}{\sqrt{A}}\right)\mathbb{E}[Q^{\alpha-1}], \end{aligned}$$

then the conditional density of U given E is

$$\begin{aligned} f_{U|E}(u|\varepsilon) &= \frac{f_{U,E}(u,\varepsilon)}{f_E(\varepsilon)} \\ &= \frac{u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)}{\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du} \\ &= \frac{u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)}{\Phi\left(\frac{B}{\sqrt{A}}\right)\mathbb{E}[Q^{\alpha-1}]}, \end{aligned} \tag{C.9}$$

with expected value given by

$$\mathbb{E}[U|E] = \int_0^\infty u f_{U|E}(u|\varepsilon)du$$

$$
\begin{aligned}
&= \frac{\int_0^\infty u^\alpha \sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du}{\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du} \\
&= \frac{\mathbb{E}[Q^\alpha]}{\mathbb{E}[Q^{\alpha-1}]}.
\end{aligned} \tag{C.10}
$$

The last two equalities follow from equation (C.18) where $Q \sim N^+\left(\frac{B}{A}, \frac{1}{\sqrt{A}}\right)$. Thus $\mathbb{E}\left[Q^\alpha\right]$ and $\mathbb{E}\left[Q^{\alpha-1}\right]$ are fractional moments of the nonnegative truncated normal distribution of Q. The expected value of a function g of U given E can be calculated similarly as

$$
\begin{aligned}
\mathbb{E}[g(U)|E] &= \int_0^\infty g(u)u^{\alpha-1}f_{U|E}(u|\varepsilon)du \\
&= \frac{\int_0^\infty g(u)u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du}{\int_0^\infty u^{\alpha-1}\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)du} \\
&= \frac{\mathbb{E}[g(Q)Q^{\alpha-1}]}{\mathbb{E}[Q^{\alpha-1}]}.
\end{aligned} \tag{C.11}
$$

When $\alpha = 1$, the conditional density is

$$
f_{U|E}(u|\varepsilon) = \frac{\sqrt{A}\phi\left(\frac{u-B/A}{1/\sqrt{A}}\right)}{\Phi\left(\frac{B}{\sqrt{A}}\right)}. \tag{C.12}
$$

Equation (C.19) gives the expected value of a truncated normal random variable. Using this equation for $Q \sim N^+\left(\frac{B}{A}, \frac{1}{\sqrt{A}}\right)$, the conditional expectation of U given E when $\alpha = 1$ is

$$
\begin{aligned}
\mathbb{E}[U|E] &= \mathbb{E}[Q] \\
&= \frac{B}{A} + \frac{\frac{1}{\sqrt{A}}\phi\left(-\frac{B/A}{1/\sqrt{A}}\right)}{1-\Phi\left(-\frac{B/A}{1/\sqrt{A}}\right)} \\
&= \frac{1}{\sqrt{A}}\left[\frac{B}{\sqrt{A}} + h\left(-\frac{B}{\sqrt{A}}\right)\right],
\end{aligned} \tag{C.13}
$$

where $h(\cdot)$ is the normal hazard function. The mode of the conditional density of U given E is located at the local maximum of $f_{U|E}(u|\varepsilon)$. The conditional density given in equation (C.12) is maximised when $\dfrac{u - B/A}{1/\sqrt{A}} = 0$, that is when $u = B/A$, therefore the conditional mode is

$$M(U|E) \;=\; \frac{B}{A}. \tag{C.14}$$

The mode has an appealing interpretation as a maximum likelihood estimator.

C.5 Truncated Normal Distributions

The probability density function of a truncated normally distributed random variable X is given in Johnson & Kotz (1970) as

$$\begin{aligned} f_X(x;\mu,\sigma) \;&=\; \frac{\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right\}}{\dfrac{1}{\sqrt{2\pi}\sigma}\displaystyle\int_A^B \exp\left\{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right\}dt} \\ &=\; \frac{\dfrac{1}{\sigma}\phi\left(\dfrac{x-\mu}{\sigma}\right)}{\Phi\left(\dfrac{B-\mu}{\sigma}\right) - \Phi\left(\dfrac{A-\mu}{\sigma}\right)}, \qquad A \le X \le B. \end{aligned} \tag{C.15}$$

The lower and upper truncation points are A and B respectively. The distribution is doubly truncated if $-\infty < A < B < \infty$. If $A = -\infty$, the distribution is singly truncated from above. If $B = \infty$, the distribution is singly truncated from below. A half normal distribution arises when $A = \mu$ and $B = \infty$ and so it is a singly truncated normal distribution, truncated from below at μ. The notation $X \sim N^+(\mu, \sigma^2)$ indicates that X has a normal distribution, with mean μ and variance σ^2, which is truncated from below at $X = 0$, i.e. $X \ge 0$.

The expected value and variance of X can be easily obtained using the moment generating function of X

$$M_X(t) \;=\; \mathbb{E}\left[e^{tX}\right]$$

$$
\begin{aligned}
&= \frac{\int_A^B e^{tx}\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx}{\Phi\left(\frac{B-\mu}{\sigma}\right)-\Phi\left(\frac{A-\mu}{\sigma}\right)} \\
&= \exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}\left[\frac{\Phi\left(\frac{B-\mu}{\sigma}-\sigma t\right)-\Phi\left(\frac{A-\mu}{\sigma}-\sigma t\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right)-\Phi\left(\frac{A-\mu}{\sigma}\right)}\right].
\end{aligned}
$$

The last equality follows from

$$
\begin{aligned}
&\int_A^B \exp\{tx\}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}dx \\
&= \int_A^B \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{tx-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}dx \\
&= \int_A^B \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}\left[-2\sigma^2tx+x^2-2x\mu+\mu^2\right]\right\}dx \\
&= \int_A^B \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{\left[x-(\mu+\sigma^2t)\right]^2}{2\sigma^2}-\frac{\mu^2-(\mu+\sigma^2t)^2}{2\sigma^2}\right\}dx \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[\mu^2-(\mu+\sigma^2t)^2\right]\right\}\int_A^B \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\tilde{\mu}}{\sigma}\right)^2\right\}dx \\
&= \exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}\int_A^B \frac{1}{\sigma}\phi\left(\frac{x-\tilde{\mu}}{\sigma}\right)dx \\
&= \exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}\left[\Phi\left(\frac{B-\tilde{\mu}}{\sigma}\right)-\Phi\left(\frac{A-\tilde{\mu}}{\sigma}\right)\right] \\
&= \exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}\left[\Phi\left(\frac{B-\mu}{\sigma}-\sigma t\right)-\Phi\left(\frac{A-\mu}{\sigma}-\sigma t\right)\right],
\end{aligned}
$$

where $\tilde{\mu}=\mu+\sigma^2t$. The n-th moment is given by

$$
\mathbb{E}[X^n]=M_X^{(n)}(0)=\left.\frac{d^nM_X(t)}{dt^n}\right|_{t=0}.
$$

The first-order derivative of the moment generating function is

$$
\begin{aligned}
M_X^{(1)}(t) &= \left(\mu+\sigma^2t\right)\exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}\left[\frac{\Phi(\alpha_B-\sigma t)-\Phi(\alpha_A-\sigma t)}{\Phi(\alpha_B)-\Phi(\alpha_A)}\right] \\
&\quad -\sigma\exp\left\{\mu t+\frac{1}{2}\sigma^2t^2\right\}\left[\frac{\phi(\alpha_B-\sigma t)-\phi(\alpha_A-\sigma t)}{\Phi(\alpha_B)-\Phi(\alpha_A)}\right],
\end{aligned}
$$

where $\alpha_A = \dfrac{A-\mu}{\sigma}$ and $\alpha_B = \dfrac{B-\mu}{\sigma}$. Evaluating this derivative at $t = 0$ gives the first moment, or expected value, as

$$\mathbb{E}[X] = M_X^{(1)}(0) = \mu - \frac{\phi(\alpha_B) - \phi(\alpha_A)}{\Phi(\alpha_B) - \Phi(\alpha_A)}\sigma. \tag{C.16}$$

The second-order derivative of the moment generating is

$$\begin{aligned} M_X^{(2)}(t) = {} & [\sigma^2 + (\mu + \sigma^2 t)^2] \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\} \left[\frac{\Phi(\alpha_B - \sigma t) - \Phi(\alpha_A - \sigma t)}{\Phi(\alpha_B) - \Phi(\alpha_A)}\right] \\ & -2\sigma(\mu + \sigma^2 t) \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\} \left[\frac{\phi(\alpha_B - \sigma t) - \phi(\alpha_A - \sigma t)}{\Phi(\alpha_B) - \Phi(\alpha_A)}\right] \\ & -\sigma^2 \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\} \times \\ & \left[\frac{(\alpha_B - \sigma t)\phi(\alpha_B - \sigma t) - (\alpha_A - \sigma t)\phi(\alpha_A - \sigma t)}{\Phi(\alpha_B) - \Phi(\alpha_A)}\right]. \end{aligned}$$

Evaluating this derivative at $t = 0$ gives the second moment as

$$\mathbb{E}\left[X^2\right] = M_X^{(2)}(0) = \sigma^2 + \mu^2 - 2\mu\sigma\frac{\phi(\alpha_B) - \phi(\alpha_A)}{\Phi(\alpha_B) - \Phi(\alpha_A)} - \sigma^2\frac{\alpha_B\phi(\alpha_B) - \alpha_A\phi(\alpha_A)}{\Phi(\alpha_B) - \Phi(\alpha_A)}. \tag{C.17}$$

Higher-order moments can be derived in a similar fashion. An alternative, but equivalent, formula for the n-th moment is

$$\mathbb{E}\left[X^n\right] = \int_A^B x^n f_X(x; \mu, \sigma)dx = \frac{\displaystyle\int_A^B x^n \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx}{\Phi\left(\dfrac{B-\mu}{\sigma}\right) - \Phi\left(\dfrac{A-\mu}{\sigma}\right)}. \tag{C.18}$$

Using the first and second moments, the variance of X is

$$\begin{aligned} Var(X) &= \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 \\ &= \left\{1 - \frac{\alpha_B\phi(\alpha_B) - \alpha_A\phi(\alpha_A)}{\Phi(\alpha_B) - \Phi(\alpha_A)} - \left[\frac{\phi(\alpha_B) - \phi(\alpha_A)}{\Phi(\alpha_B) - \Phi(\alpha_A)}\right]^2\right\}\sigma^2. \end{aligned}$$

Substituting the values of α_A and α_B into equations (C.16) and (C.17) gives the expected value and variance of X as

$$\mathbb{E}[X] = \mu - \frac{\phi\left(\dfrac{B-\mu}{\sigma}\right) - \phi\left(\dfrac{A-\mu}{\sigma}\right)}{\Phi\left(\dfrac{B-\mu}{\sigma}\right) - \Phi\left(\dfrac{A-\mu}{\sigma}\right)}\sigma, \tag{C.19}$$

$$Var(X) = \left\{ 1 - \frac{\left(\frac{B-\mu}{\sigma}\right)\phi\left(\frac{B-\mu}{\sigma}\right) - \left(\frac{A-\mu}{\sigma}\right)\phi\left(\frac{A-\mu}{\sigma}\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)} - \left[\frac{\phi\left(\frac{B-\mu}{\sigma}\right) - \phi\left(\frac{A-\mu}{\sigma}\right)}{\Phi\left(\frac{B-\mu}{\sigma}\right) - \Phi\left(\frac{A-\mu}{\sigma}\right)}\right]^2 \right\} \sigma^2.$$

These agree with the formulae for the expected value and variance of X given in Johnson & Kotz (1970).

C.6 Hazard Functions

The hazard function is the ratio of the probability density function $f(x)$ to the survival function $S(x)$.

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)},$$

where $S(x) = 1 - F(x)$ and $F(x)$ is the cumulative distribution function. The formula for the hazard function of the normal distribution is

$$h(x) = \frac{\phi(x)}{1 - \Phi(x)},$$

where $\phi(x)$ is the probability density function of the standard normal distribution and $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

The derivative of the hazard function with respect to x is

$$h'(x) = h(x)\left[\frac{f'(x)}{f(x)} + h(x)\right],$$

and for the normal hazard function, the derivative simplifies to

$$h'(x) = h(x)[-x + h(x)].$$

Elandt-Johnson & Johnson (1980) provide further details on hazard functions and their use in the analysis of survival models.

C.7 Taylor Approximations

Let X be a random variable with realisation x. Also let $\mu_x = \mathbb{E}[X]$ and $\sigma_x^2 = Var(X)$. A smooth function $f(x)$ can be approximated using a Taylor polynomial centred around its mean. The Taylor series expansion of $f(x)$ about the point μ_x is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(\mu_x)}{n!}(x-\mu_x)^n,$$

where $f^{(n)}(x)$ is the n-th order derivative of $f(x)$.

Let the Taylor approximation of $f(x)$ be denoted by $\widehat{f}(x)$. The first-order Taylor series expansion of $f(x)$ and its expected value are

$$\widehat{f}(x) = f(\mu_x) + (x-\mu_x) f'(\mu_x), \tag{C.20}$$

$$\begin{aligned} \mathbb{E}\left[\widehat{f}(x)\right] &= f(\mu_x) + (\mu_x - \mu_x) f'(\mu_x) \\ &= f(\mu_x), \end{aligned} \tag{C.21}$$

where

$$f'(\mu_x) = \left.\frac{\partial f(x)}{\partial x}\right|_{x=\mu_x}.$$

The second-order Taylor series expansion of $f(x)$ and its expected value are

$$\widehat{f}(x) = f(\mu_x) + (x-\mu_x) f'(\mu_x) + \frac{1}{2}(x-\mu_x)^2 f''(\mu_x),$$

$$\begin{aligned} \mathbb{E}\left[\widehat{f}(x)\right] &= f(\mu_x) + (\mu_x - \mu_x) f'(\mu_x) + \frac{1}{2}\mathbb{E}\left[(x-\mu_x)^2\right] f''(\mu_x) \\ &= f(\mu_x) + \frac{\sigma_x^2}{2} f''(\mu_x), \end{aligned}$$

where

$$f''(\mu_x) = \left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{x=\mu_x}.$$

Stewart (1995) gives a summary of Taylor series.

Appendix D

Information Matrices

D.1 Log-likelihood Function

The likelihood function of $\boldsymbol{\theta}$ is the joint probability density function of N random variables $Y_1, \ldots, Y_N$ conditional on $\boldsymbol{\theta}$. For a sample of N independent observations the likelihood function is given by

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) &= f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}) \\ &= f_{Y_1,\ldots,Y_N}(y_1, \ldots, y_N; \boldsymbol{\theta}) \\ &= \prod_{i=1}^{N} f_{Y_i}(y_i; \boldsymbol{\theta}), \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ is a vector of k parameters that require estimation and $f_{Y_i}(y_i; \boldsymbol{\theta})$ is the probability density function of random variable Y_i. Taking a logarithmic transformation gives the log-likelihood function

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{N} \ln f_{Y_i}(y_i; \boldsymbol{\theta}).$$

For the statistical model

$$Y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + E_i, \tag{D.1}$$

the probability density function of Y_i, derived using a transformation of variables, is

$$f_{Y_i}(y_i) \;=\; f_{E_i}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})),$$

where $f_{Y_i}(y_i)$ is the density of Y_i evaluated at y_i and $f_{E_i}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}))$ is the density of E_i evaluated at $\varepsilon_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})$. Under model (D.1), the likelihood function is

$$\mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; \prod_{i=1}^{N} f_{E_i}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}); \boldsymbol{\theta}),$$

and the log-likelihood function is

$$\ln \mathscr{L}(\boldsymbol{\theta}; \boldsymbol{y}) \;=\; \sum_{i=1}^{N} \ln f_{E_i}(y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta}); \boldsymbol{\theta}).$$

D.2 Information Matrix for a Single Observation

The per observation expected Fisher information matrix of $\boldsymbol{\theta}$ for the i-th observation y_i taken at $\boldsymbol{x}_i$ is given by the $k \times k$ symmetric matrix

$$I_i(\boldsymbol{\theta}) \;=\; Cov\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right), \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right] \;=\; \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}\right)^T\right], \tag{D.2}$$

where $f_{Y_i} = f_{Y_i}(y_i; \boldsymbol{\theta})$ is the probability density function of random variable Y_i. The likelihood for a single observation is just the density function

$$\mathscr{L}(\boldsymbol{\theta}; y_i) \;=\; f_{Y_i}(y_i; \boldsymbol{\theta}),$$

hence the partial derivative $\dfrac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\theta}}$ is the score (of the log-likelihood). The expectation of the score is zero, hence the per observation expected Fisher information matrix is just the covariance of the score. The (j, l)-th element is given by

$$I_i(\boldsymbol{\theta})_{(j,l)} \;=\; Cov\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \theta_j}\right), \left(\frac{\partial \ln f_{Y_i}}{\partial \theta_l}\right)\right] \;=\; \mathbb{E}\left[\left(\frac{\partial \ln f_{Y_i}}{\partial \theta_j}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \theta_l}\right)\right].$$

Under certain regularity conditions

$$I_i(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right], \tag{D.3}$$

i.e. the (j,l)-th element is given by

$$I_i(\boldsymbol{\theta})_{(j,l)} \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln f_{Y_i}}{\partial\theta_j\partial\theta_l}\right].$$

D.3 Information Matrix for N Observations

The expected Fisher information matrix of $\boldsymbol{\theta}$ for N observations $y_1,\ldots,y_N$ taken at $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$ is given by the $k\times k$ symmetric matrix

$$I_N(\boldsymbol{\theta}) \;=\; Cov\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\boldsymbol{\theta}}\right),\left(\frac{\partial \ln \mathscr{L}}{\partial\boldsymbol{\theta}}\right)^T\right] \;=\; \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \ln \mathscr{L}}{\partial\boldsymbol{\theta}}\right)^T\right],$$

where $\mathscr{L}=\mathscr{L}(\boldsymbol{\theta};\boldsymbol{y})$ is the likelihood function for the N observations. The partial derivative $\dfrac{\partial \ln \mathscr{L}}{\partial\boldsymbol{\theta}}$ is known as the score (of the log-likelihood). The expectation of the score is zero, hence the expected Fisher information matrix is just the covariance of the score. The (j,l)-th element of the Fisher information matrix is given by

$$I_N(\boldsymbol{\theta})_{(j,l)} \;=\; Cov\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\theta_j}\right),\left(\frac{\partial \ln \mathscr{L}}{\partial\theta_l}\right)\right] \;=\; \mathbb{E}\left[\left(\frac{\partial \ln \mathscr{L}}{\partial\theta_j}\right)\left(\frac{\partial \ln \mathscr{L}}{\partial\theta_l}\right)\right].$$

Under certain regularity conditions

$$I_N(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln \mathscr{L}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right],$$

i.e. the (j,l)-th element is given by

$$I_N(\boldsymbol{\theta})_{(j,l)} \;=\; -\mathbb{E}\left[\frac{\partial^2 \ln \mathscr{L}}{\partial\theta_j\partial\theta_l}\right].$$

If $\hat{\boldsymbol{\theta}}$ is an estimator for $\boldsymbol{\theta}$, then the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be obtained by inverting the expected Fisher information matrix of $\boldsymbol{\theta}$, i.e.

$$Cov(\hat{\boldsymbol{\theta}}) \;=\; I_N(\boldsymbol{\theta})^{-1}.$$

It is also worth noting two useful properties of the expected Fisher information matrix. Firstly, the expected Fisher information for a sample of N independent observations is equal to N times the Fisher information for a single observation, i.e.

$$I_N(\boldsymbol{\theta}) \;=\; NI_i(\boldsymbol{\theta}).$$

Secondly, it is dependent on the choice of parameterisation. Suppose the parameter $\boldsymbol{\theta}$ is reparameterised to $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ with $\eta_j = g_j(\boldsymbol{\theta})$ where each g_j is one-to-one so its inverse $g_j^{-1}(\boldsymbol{\eta}) = \theta_j$ exists. The Fisher information $I_N^*(\boldsymbol{\eta})$ for the new parameterisation is obtained using the chain rule and is given by Schervish (1995) as

$$I_N^*(\boldsymbol{\eta}) \;=\; J(\boldsymbol{\eta})^T I_N(\boldsymbol{\theta}(\boldsymbol{\eta}))J(\boldsymbol{\eta}),$$

where $J(\boldsymbol{\eta})$ is the Jacobian matrix with elements

$$J(\boldsymbol{\eta})_{(j,l)} \;=\; \frac{\partial g_j^{-1}(\boldsymbol{\eta})}{\partial \eta_l}, \qquad (j, l = 1, \ldots, k)$$

and $\theta(\boldsymbol{\eta}) = (g_1^{-1}(\boldsymbol{\eta}), \ldots, g_k^{-1}(\boldsymbol{\eta}))$.

D.4 Partitioned Information Matrix

Let $\boldsymbol{\beta}$ be a vector of p parameters and let $\boldsymbol{\tau}$ be a vector of $k - p$ parameters. If $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ then $\boldsymbol{\theta}$ is a k-dimensional parameter vector with partitioned per observation expected Fisher information matrix

$$I_i(\boldsymbol{\theta}) \;=\; \mathbb{E}\left[\begin{array}{c|c} \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}}\right)^T \\ \hline \left\{\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}}\right)^T\right\}^T & \left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}}\right)\left(\frac{\partial \ln f_{Y_i}}{\partial \boldsymbol{\tau}}\right)^T \end{array}\right]. \tag{D.4}$$

Under certain regularity conditions the expected Fisher information matrix can be written equivalently as

$$I_i(\boldsymbol{\theta}) \;=\; -\mathbb{E}\left[\begin{array}{c|c} \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\tau}^T} \\ \hline \left(\dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\tau}^T}\right)^T & \dfrac{\partial^2 \ln f_{Y_i}}{\partial\boldsymbol{\tau}\partial\boldsymbol{\tau}^T} \end{array}\right]. \tag{D.5}$$

D.5 Eigendecomposition of Partitioned Information Matrices

Suppose the partitioned per observation expected Fisher information matrix (D.4) or (D.5) has the form

$$I_i(\boldsymbol{\theta}) \;=\; \left[\begin{array}{c|c} f_\beta \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}_{\beta,\tau}^T \\ \hline \boldsymbol{f}_{\beta,\tau}\boldsymbol{f}^T(\boldsymbol{x}_i) & F_\tau \end{array}\right], \tag{D.6}$$

where f_β is a scalar-valued function, $\boldsymbol{f}(\boldsymbol{x}_i)$ is a vector of length p, $\boldsymbol{f}_{\beta,\tau}$ is a vector-valued function of length $k-p$ and F_τ is a $(k-p)$-dimensional square matrix. An information matrix of this form may arise from a linear model

$$Y_i \;=\; \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{\beta} + E_i, \qquad i = 1, \ldots, N,$$

where a distributional assumption is placed on the E_i with the distribution having parameters $\boldsymbol{\tau}$. Thus the full parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{k-p})$.

Let $\boldsymbol{f}_\theta(\boldsymbol{x}_i)$ be the $k \times 1$ vector with

$$\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) \;=\; \left[\begin{array}{c|c} \sqrt{f_\beta}\boldsymbol{f}^T(\boldsymbol{x}_i) & \dfrac{1}{\sqrt{f_\beta}}\boldsymbol{f}_{\beta,\tau}^T \end{array}\right],$$

where $\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$ is the non-Hermitian or non-conjugate transpose of $\boldsymbol{f}_\theta(\boldsymbol{x}_i)$[1]. The symmetric $k \times k$ matrix $\boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i)$ is

$$\boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) \;=\; \left[\begin{array}{c|c} f_\beta \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}_{\beta,\tau}^T \\ \hline \boldsymbol{f}_{\beta,\tau}\boldsymbol{f}^T(\boldsymbol{x}_i) & \dfrac{1}{f_\beta}\boldsymbol{f}_{\beta,\tau}\boldsymbol{f}_{\beta,\tau}^T \end{array}\right].$$

If $F_\tau = (1/f_\beta)\boldsymbol{f}_{\beta,\tau}\boldsymbol{f}_{\beta,\tau}^T$ then information matrix (D.6) can be expressed as

$$I_i(\boldsymbol{\theta}) \;=\; \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i).$$

For continuous design (5.8) with n distinct design points, if the information matrix $M(\xi)$ is a weighted sum of per observation expected Fisher information matrices, then

$$\begin{aligned} M(\xi) &= \sum_{i=1}^{n} w_i I_i(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} w_i \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) \\ &= F^T W F, \end{aligned}$$

where $\sum_{i=1}^{n} w_i = 1$ and

$$\begin{aligned} F^T &= \left[\boldsymbol{f}_\theta(\boldsymbol{x}_1), \ldots, \boldsymbol{f}_\theta(\boldsymbol{x}_n)\right], \\ W &= \operatorname{diag}\left(w_1, \ldots, w_n\right). \end{aligned}$$

Usually interest is in estimating the $\boldsymbol{\beta}$ parameters. However, when considering the extended full parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$, the matrix F_τ will not necessarily equal $(1/f_\beta)\boldsymbol{f}_{\beta,\tau}\boldsymbol{f}_{\beta,\tau}^T$. This situation occurs for example in the case of simple linear regression which has per observation expected Fisher information matrix

$$I_i(\boldsymbol{\theta}) \;=\; \left[\begin{array}{c|c} \dfrac{1}{\sigma^2}\boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) & \mathbf{0} \\ \hline \mathbf{0} & \dfrac{1}{2\sigma^4} \end{array}\right].$$

[1]The complex conjugate transpose of $\boldsymbol{x}$ is usually denoted $\boldsymbol{x}^T$ whereas the non-conjugate or non-Hermitian transpose is usually denoted $\boldsymbol{x}.^T$. For $\boldsymbol{x} = \sqrt{-c}$, using the complex conjugate transpose gives $\boldsymbol{x}\boldsymbol{x}^T = c$ while the non-conjugate transpose gives $\boldsymbol{x}\boldsymbol{x}.^T = -c$.

It may also occur when the per observation expected Fisher information matrix is approximated using Method 2 or 3 of Chapter 3.

Let $\mathscr{C}$ be a symmetric $k \times k$ 'correction matrix' given by

$$\mathscr{C} = \left[\begin{array}{c|c} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times(k-p)} \\ \hline \mathbf{0}_{(k-p)\times p} & F_\tau - \frac{1}{f_\beta}\boldsymbol{f}_{\beta,\tau}\boldsymbol{f}_{\beta,\tau}^T \end{array}\right],$$

then information matrix (D.6) becomes

$$I_i(\boldsymbol{\theta}) = \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C}. \tag{D.7}$$

The eigenvalue decomposition of $\mathscr{C}$ is

$$\mathscr{C} = Q_k\Lambda_k Q_k^T = \textstyle\sum_{j=1}^k \lambda_j \boldsymbol{q}_j\boldsymbol{q}_j^T,$$

where

$$\begin{aligned} Q_k &= [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k] \\ &= [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p, \boldsymbol{q}_{p+1}, \ldots, \boldsymbol{q}_k], \end{aligned}$$

$$\begin{aligned} \Lambda_k &= \operatorname{diag}(\lambda_1, \ldots, \lambda_k) \\ &= \operatorname{diag}\left(\mathbf{0}_p^T, \lambda_{p+1}, \ldots, \lambda_k\right). \end{aligned}$$

The λ_j and $\boldsymbol{q}_j$ are the eigenvalues and eigenvectors of $\mathscr{C}$ respectively and $\boldsymbol{e}_j$ is a $k \times 1$ coordinate or unit vector with a 1 in the j-th position and zeros elsewhere. Because the last $k - p$ columns of $\mathscr{C}$ are the only linearly independent columns, $\lambda_{p+1}, \ldots, \lambda_k$ are the only nonzero eigenvalues.

The structure of $\mathscr{C}$ can be exploited to improve computational efficiency. Let C_{22} be the nonzero submatrix of $\mathscr{C}$,

$$C_{22} = F_\tau - \frac{1}{f_\beta}\boldsymbol{f}_{\beta,\tau}\boldsymbol{f}_{\beta,\tau}^T.$$

If the eigenvalue decomposition of C_{22} is

$$C_{22} = Z\Lambda_{22}Z^T = \sum_{j=p+1}^{k} \lambda_j \boldsymbol{z}_j \boldsymbol{z}_j^T,$$

where

$$\begin{aligned} Z &= [\boldsymbol{z}_{p+1}, \ldots, \boldsymbol{z}_k], \\ \Lambda_{22} &= \operatorname{diag}(\lambda_{p+1}, \ldots, \lambda_k), \end{aligned}$$

and λ_j and $\boldsymbol{z}_j$ are the respective eigenvalues and eigenvectors of C_{22} then

$$\boldsymbol{q}_j^T = \left[\mathbf{0}_p^T, \boldsymbol{z}_j^T\right], \qquad j = p+1, \ldots, k,$$

and

$$\Lambda_k = \operatorname{diag}\left(\mathbf{0}_p^T, \operatorname{diag}(\Lambda_{22})\right) = \left[\begin{array}{c|c} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times(k-p)} \\ \hline \mathbf{0}_{(k-p)\times p} & \Lambda_{22} \end{array}\right].$$

Letting

$$Q_{22} = [\boldsymbol{q}_{p+1}, \ldots, \boldsymbol{q}_k] = \left[\begin{array}{c} \mathbf{0}_{p\times(k-p)} \\ \hline Z \end{array}\right],$$

then

$$Q_k = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p, Q_{22}] = \left[\begin{array}{c|c} I_p & \mathbf{0}_{p\times(k-p)} \\ \hline \mathbf{0}_{(k-p)\times p} & Z \end{array}\right],$$

and it follows that

$$\begin{aligned} \mathscr{C} &= Q_k \Lambda_k Q_k^T \\ &= Q_{22}\Lambda_{22}Q_{22}^T \\ &= \sum_{j=p+1}^{k} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j^T. \end{aligned} \tag{D.8}$$

Finally, substituting equation (D.8) into equation (D.7) gives the per observation expected Fisher information matrix as

$$I_i(\boldsymbol{\theta}) = \boldsymbol{f}_\theta(\boldsymbol{x}_i)\boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \sum_{j=p+1}^{k} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j^T.$$

If the information matrix $M(\xi)$ is a weighted sum of per observation expected Fisher information matrices, the above eigenvalue decomposition gives

$$\begin{aligned} M(\xi) &= \sum_{i=1}^{n} w_i I_i(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} w_i \left\{ \boldsymbol{f}_\theta(\boldsymbol{x}_i) \boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C} \right\} \\ &= \sum_{i=1}^{n} w_i \boldsymbol{f}_\theta(\boldsymbol{x}_i) \boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \mathscr{C} \\ &= \sum_{i=1}^{n} w_i \boldsymbol{f}_\theta(\boldsymbol{x}_i) \boldsymbol{f}_\theta^T(\boldsymbol{x}_i) + \sum_{j=p+1}^{k} \lambda_j \boldsymbol{q}_j \boldsymbol{q}_j^T \\ &= F^T W F + Q_{22} \Lambda_{22} Q_{22}^T. \end{aligned}$$

Let

$$\begin{aligned} R^T &= [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n, \boldsymbol{r}_{n+1}, \ldots, \boldsymbol{r}_{n+k-p}] \\ &= [\boldsymbol{f}_\theta(\boldsymbol{x}_1), \ldots, \boldsymbol{f}_\theta(\boldsymbol{x}_n), \boldsymbol{q}_{p+1}, \ldots, \boldsymbol{q}_k] \\ &= [F, Q_{22}], \end{aligned}$$

$$\begin{aligned} S &= \operatorname{diag}(s_1, \ldots, s_n, s_{n+1}, \ldots, s_{n+k-p}) \\ &= \operatorname{diag}(w_1, \ldots, w_n, \lambda_{p+1}, \ldots, \lambda_k) \\ &= \operatorname{diag}(\operatorname{diag}(W), \operatorname{diag}(\Lambda_{22})), \end{aligned}$$

then

$$M(\xi) = R^T S R = \textstyle\sum_{i=1}^{n+k-p} s_i \boldsymbol{r}_i \boldsymbol{r}_i^T.$$

This shows that, although the per observation expected Fisher information matrix $I_i(\boldsymbol{\theta})$ may not always be expressible as a column vector multiplied by its own transpose, the information matrix $M(\xi)$ can be expressed in such a manner using an eigenvalue decomposition and a little algebra.

If $k - p = 2$ then the eigenvalues of $\mathscr{C}$ are

$$\text{diag}(\Lambda_{22}) = \begin{bmatrix} \lambda_{k-1} \\ \lambda_k \end{bmatrix} = \frac{1}{2g_1} \begin{bmatrix} g_1 + g_2 + \sqrt{g_3} \\ g_1 + g_2 - \sqrt{g_3}, \end{bmatrix}$$

and the eigenvectors $Q_{22} = [\boldsymbol{q}_{k-1}, \boldsymbol{q}_k]$ are

$$\boldsymbol{q}_{k-1} = \frac{|g_4|}{g_4 \left(|g_1 - g_2 - \sqrt{g_3}|^2 + 4|g_4|^2\right)^{1/2}} \begin{bmatrix} g_1 - g_2 - \sqrt{g_3} \\ 2g_4 \end{bmatrix},$$

$$\boldsymbol{q}_k = \frac{|g_4|}{g_4 \left(|g_1 - g_2 + \sqrt{g_3}|^2 + 4|g_4|^2\right)^{1/2}} \begin{bmatrix} g_1 - g_2 + \sqrt{g_3} \\ 2g_4 \end{bmatrix},$$

where

$$\begin{aligned}
g_1 &= f_\beta F_\tau(2,2) - \boldsymbol{f}_{\beta,\tau}(2)^2, \\
g_2 &= f_\beta F_\tau(1,1) - \boldsymbol{f}_{\beta,\tau}(1)^2, \\
g_3 &= f_\beta^2 \left\{F_\tau(2,2) - F_\tau(1,1)\right\}^2 + 4 f_\beta F_\tau(1,2)^2 + 2 f_\beta \boldsymbol{f}_{\beta,\tau}(1)^2 \left\{F_\tau(2,2) - F_\tau(1,1)\right\} \\
&\quad -2 f_\beta \boldsymbol{f}_{\beta,\tau}(2)^2 \left\{F_\tau(2,2) - F_\tau(1,1)\right\} - 8 f_\beta \boldsymbol{f}_{\beta,\tau}(1) \boldsymbol{f}_{\beta,\tau}(2) F_\tau(1,2) \\
&\quad + \left\{\boldsymbol{f}_{\beta,\tau}(1)^2 + \boldsymbol{f}_{\beta,\tau}(2)^2\right\}^2 \\
g_4 &= -f_\beta F_\tau(1,2) + \boldsymbol{f}_{\beta,\tau}(1) \boldsymbol{f}_{\beta,\tau}(2).
\end{aligned}$$

The notation $F_\tau(i,j)$ refers to the (i,j)-th element of the matrix F_τ and $\boldsymbol{f}_{\beta,\tau}(i)$ refers to the i-th element of the vector $\boldsymbol{f}_{\beta,\tau}$.

Appendix E

Matrix Inverses

E.1 Inverse of a Partitioned Matrix

Suppose we wish to invert the partitioned $k \times k$ matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with A_{11} and A_{22} both square and with respective sizes $p \times p$ and $(k-p) \times (k-p)$ say. Suppose we partition the inverse in the same way and write

$$A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}.$$

Then we can treat the submatrices as if they were elements and derive

$$\begin{aligned} A^{11} &= \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}, \\ A^{12} &= -A^{11}A_{12}A_{22}^{-1}, \\ A^{21} &= -A_{22}^{-1}A_{21}A^{11}, \\ A^{22} &= A_{22}^{-1} - A^{21}A_{12}A_{22}^{-1}, \end{aligned}$$

assuming that A_{22} and $\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)$ are nonsingular. The matrix $\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)$ is the *Schur complement* of the block A_{22}. This matrix in-

verse is given in Healy (2000). Alternatively, the blocks of the inverse may be expressed as

$$\begin{aligned} A^{11} &= A_{11}^{-1} - A^{12}A_{21}A_{11}^{-1}, \\ A^{12} &= -A_{11}^{-1}A_{12}A^{22}, \\ A^{21} &= -A^{22}A_{21}A_{11}^{-1}, \\ A^{22} &= \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}, \end{aligned}$$

assuming that A_{11} and its Schur complement $\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)$ are nonsingular (Watt 2006).

E.2 Inverse of a Sum of Two Matrices

If A is positive definite and symmetric and

$$B = A + d\boldsymbol{x}\boldsymbol{x}^T,$$

with d a scalar, $\boldsymbol{x}$ a vector and B positive definite (e.g. if $d > 0$) then Anderson (1984) gives the determinant and inverse of B respectively as

$$|B| = \left(1 + d\boldsymbol{x}^T A^{-1}\boldsymbol{x}\right)|A|, \tag{E.1}$$

$$B^{-1} = A^{-1} - \frac{d}{1 + d\boldsymbol{x}^T A^{-1}\boldsymbol{x}} A^{-1}\boldsymbol{x}\boldsymbol{x}^T A^{-1}. \tag{E.2}$$

If a nonsingular matrix M can be obtained in the iterative manner

$$M_{(k)} = M_{(k-1)} + d\boldsymbol{x}\boldsymbol{x}^T,$$

where $M_{(k)}$ is the matrix M at the k-th iteration, then equations (E.1) and (E.2) can be used iteratively to find the determinant and inverse of M at the k-th iteration. They are given respectively by

$$|M_{(k)}| = \left(1 + d\boldsymbol{x}^T M_{(k-1)}^{-1}\boldsymbol{x}\right)|M_{(k-1)}|,$$

$$M_{(k)}^{-1} = M_{(k-1)}^{-1} - \frac{d}{1 + d\boldsymbol{x}^T M_{(k-1)}^{-1} \boldsymbol{x}} M_{(k-1)}^{-1} \boldsymbol{x}\boldsymbol{x}^T M_{(k-1)}^{-1}.$$

Appendix F

Optimum Design

F.1 Pseudocode for Torsney's Multiplicative Algorithm

Algorithm (5.13) provides a method for finding optimising distributions for optimum experimental designs. More specifically, it is a multiplicative algorithm for finding the optimal design weights for a selection of candidate design points. The candidate points are usually, but not necessarily, a grid of points over the design space $\mathscr{X}$. The weights converge to zero for any points that are not support points of the optimum design. The algorithm often gives a distribution defined on a disjoint cluster of points. Typically, within each cluster is a single true support point with nonzero weight. If the algorithm were to carry on indefinitely, the weights for the remaining points in each cluster would ideally converge to zero. Clearly limitations of time and computational resources mean that rules should be set when running the algorithm to speed up convergence.

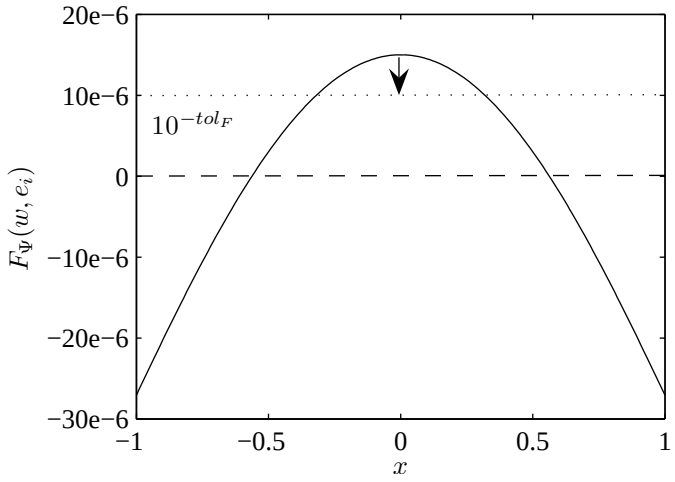

Figure F.1: Stopping criteria for algorithm (5.13).

F.1.1 Stopping criteria

The first condition of the General Equivalence Theorem (5.10), that $F_\Psi(\boldsymbol{w}^*, \boldsymbol{e}_i) = 0$ for $w_i^* > 0$, states that the derivative is zero at the optimum design points, which have nonzero weight. With regards to algorithm (5.13), the derivative will *converge* to zero at the support points. Hence a tolerance, tol_F, must be set for determining computational convergence. A stopping rule for the algorithm is thus

$$\max_{1 \le i \le n} \left\{ F_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_i) \right\} \le 10^{-tol_F}.$$

That is, the algorithm will terminate when the maximum value of the derivative $F_i^{(k)} = F_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_i)$ is close to zero, $\le 10^{-tol_F}$. Torsney (1977) suggests $tol_F = n$, however if the candidate support points are a fine grid then n will be large and $tol_F = 6$ may be suitable. If $tol_F = 6$, the algorithm will terminate when the maximum value of the derivative is less than 0.000001. In Figure F.1 the algorithm terminates when the maxima of the curve hits the dashed line where $F_i = 10^{-tol_F}$.

F.1.2 Assigning zero weights

From the second condition of the General Equivalence Theorem (5.10), that $F_\Psi(\boldsymbol{w}^*, \boldsymbol{e}_i) \leq 0$ for $w_i^* = 0$, it follows that

$$w_i^{(k)} < 10^{-tol_1} \quad \text{and} \quad F_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_i) < -10^{-tol_2} \quad \longrightarrow \quad w_i^{(k)} = 0,$$

for tolerance values $tol_1, tol_2 > 0$. That is, if the weight $w_i^{(k)}$ at the k-th iteration is small, $< 10^{-tol_1}$ say, and the derivative $F_i^{(k)}$ is large and negative, $< -10^{-tol_2}$ say, then assign a value of zero to the weight $w_i^{(k)}$. If the tolerances tol_1 or tol_2 are set too high then the algorithm will require many iterations to converge. However, if the tolerances are set too low, there may be many points within each cluster of support points, i.e. the algorithm hasn't converged satisfactorily. If tol_1 is too low then weights are more easily set to zero. Conversely, if tol_1 is too high then some weights may not be set to zero that should be set to zero. A weight should be set to zero if its derivative is not close to zero, so if tol_2 is set too low then some weights may be set to zero that should not be set to zero. Conversely, if tol_2 is too high then some weights may not be set to zero that should be set to zero. Empirical evidence suggests that $tol_1 = 4$ and $tol_2 = 4$ are suitable choices. In this case a weight will only be set to zero if the weight has a small value, $w_i^{(k)} < 0.0001$, and the magnitude of the derivative is large, $F_i^{(k)} < -0.0001$.

In Figure F.2 there should clearly be one support point at $x = 0$ with weight 1. The arrows point to candidate points that have very small weights close to zero and derivatives not close to zero. If these candidate points are in the cluster of support points at iteration k then their weight should be set to zero because they are not support points.

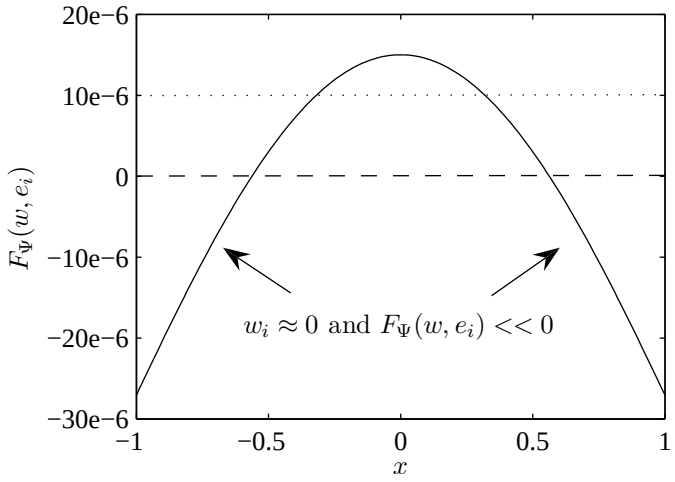

Figure F.2: Rule for assigning a value of zero to weights in algorithm (5.13).

F.1.3 Psuedocode

The procedure for finding the optimising distribution of weights using algorithm (5.13) and the above rules is as follows.

1. Initialise algorithm at $k = 0$ with $w_i^{(0)} = 1/n$ and $F_i^{(0)} = -1$.
2. Calculate $M(\xi^{(k)})$.
3. Calculate $F_i^{(k)}$.
4. If $w_i^{(k)} < 10^{-tol_1}$ and $F_i^{(k)} < -10^{-tol_2}$ then $w_i^{(k)} = 0$.
5. Ensure $\sum w_i^{(k)} = 1$ by calculating $w_i^{(k)} = \dfrac{w_i^{(k)}}{\sum w_j^{(k)}}$.
6. Calculate new weights $w_i^{(k+1)} = \dfrac{w_i^{(k)} \left\{G_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_i)\right\}^\delta}{\sum_{i=1}^n w_j^{(k)} \left\{G_\Psi(\boldsymbol{w}^{(k)}, \boldsymbol{e}_j)\right\}^\delta}$.
7. If $\max\{F_i^{(k)}\} \leq 10^{-tol_F}$ then STOP. Final weights are $w_i = w_i^{(k)}$. Else if $\max\{F_i^{(k)}\} > 10^{-tol_F}$ then $k = k + 1$ and go to step 2.

F.1.4 A check for the coded algorithm

Using definition (5.9) of the Fréchet directional derivative

$$F_\Psi(\boldsymbol{w}, \boldsymbol{e}_i) \;=\; \frac{\partial \Psi}{\partial w_i} - \sum_{j=1}^{n} \frac{\partial \Psi}{\partial w_j} w_j,$$

gives the weighted sum

$$\textstyle\sum_{j=1}^{n} w_j F_\Psi(\boldsymbol{w}, \boldsymbol{e}_j) \;=\; \displaystyle\sum_{j=1}^{n} \frac{\partial \Psi}{\partial w_j} w_j - \sum_{j=1}^{n} \frac{\partial \Psi}{\partial w_j} w_j \;=\; 0.$$

This may be helpful in verifying if the algorithm has been coded correctly.

F.2 Some Useful Matrix Properties

The following properties are useful in proving some results about optimum designs.

Theorem F.2.1 Let $\boldsymbol{f}(\boldsymbol{x}_i) = [f(x_{1i}), f(x_{2i}), \ldots, f(x_{ji}), \ldots, f(x_{pi})]^T$. For matrix $A = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)$, the vector $\boldsymbol{v} = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)$ is the j-th column of A if $\boldsymbol{f}(\boldsymbol{x}_i) = [f(x_{1i}), f(x_{2i}), \ldots, 1, \ldots, f(x_{pi})]^T$, that is, if $f(x_{ji}) = 1$.

Proof If $\boldsymbol{f}(\boldsymbol{x}_i) = [f(x_{1i}), f(x_{2i}), \ldots, f(x_{ji}), \ldots, f(x_{pi})]^T$ then A is the symmetric $p \times p$ matrix

$$A \;=\; \sum w_i \begin{bmatrix} f(x_{1i})^2 & f(x_{1i})f(x_{2i}) & \ldots & f(x_{1i})f(x_{ji}) & \ldots & f(x_{1i})f(x_{pi}) \\ & f(x_{2i})^2 & \ldots & f(x_{2i})f(x_{ji}) & \ldots & f(x_{2i})f(x_{pi}) \\ & & & \vdots & & \vdots \\ & & & f(x_{ji})^2 & \ldots & f(x_{ji})f(x_{pi}) \\ & & & & & \vdots \\ & & & & & f(x_{pi})^2 \end{bmatrix}$$

The vector $\boldsymbol{v}$ is more explicitly written

$$\boldsymbol{v} \;=\; \sum w_i [f(x_{1i}), f(x_{2i}), \ldots, f(x_{ji}), \ldots, f(x_{pi})]^T.$$

Clearly, $\boldsymbol{v}$ is the j-th column of A only if $f(x_{ji}) = 1$, that is, if the j-th element of $\boldsymbol{v} = \sum w_i$. □

Theorem F.2.2 For nonsingular $p \times p$ matrix A, if the $p \times 1$ vector $\boldsymbol{v}$ is the j-th column of A then

$$A^{-1}\boldsymbol{v} \;=\; \boldsymbol{e}_j,$$

where $\boldsymbol{e}_j$ is the coordinate vector.

Proof Let the columns of A be denoted by $\boldsymbol{a}_j$, then

$$A \;=\; [\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \ldots \quad \boldsymbol{a}_p\ ].$$

Assume that $\boldsymbol{v}$ is a column vector of A, that is

$$A \;=\; [\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \ldots \quad \boldsymbol{v} \quad \ldots \quad \boldsymbol{a}_p\ ].$$

For nonsingular A, premultiplication of A by A^{-1} gives the identity matrix

$$\begin{aligned} A^{-1}A \;=\; I_p \;&=\; [\ \boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \ldots \quad \boldsymbol{e}_j \quad \ldots \quad \boldsymbol{e}_p\ ] \\ &=\; [\ A^{-1}\boldsymbol{a}_1 \quad A^{-1}\boldsymbol{a}_2 \quad \ldots \quad A^{-1}\boldsymbol{v} \quad \ldots \quad A^{-1}\boldsymbol{a}_p\ ]. \end{aligned}$$

Therefore, if the j-th column is $\boldsymbol{v}$ then $A^{-1}\boldsymbol{v} = \boldsymbol{e}_j$. □

Corollary F.2.1 The scalar $\boldsymbol{v}^T A^{-1}\boldsymbol{v}$ is the j-th element of $\boldsymbol{v}$ since

$$\boldsymbol{v}^T A^{-1}\boldsymbol{v} \;=\; \boldsymbol{v}^T\boldsymbol{e}_j \;=\; v_j.$$

Thus if $\boldsymbol{v} = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)$ and $A = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)$ with

$$\boldsymbol{f}(\boldsymbol{x}_i) \;=\; \sum w_i [f(x_{1i}), f(x_{2i}), \ldots, 1, \ldots, f(x_{pi})]^T,$$

then

$$\boldsymbol{v}^T A^{-1}\boldsymbol{v} \;=\; \sum w_i,$$

since $\sum w_i$ is the j-th element of $\boldsymbol{v}$.

Proposition F.2.1 If $\boldsymbol{v} = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)$ and $A = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)$ with

$$\boldsymbol{f}(\boldsymbol{x}_i) = \sum w_i [f(x_{1i}), f(x_{2i}), \ldots, f(x_{ji}), \ldots, f(x_{pi})]^T,$$

that is, $f(x_{ji})$ is not necessarily a constant, then

$$\boldsymbol{v}^T A^{-1} \boldsymbol{v} \;=\; \sum w_i,$$

although $A^{-1}\boldsymbol{v} \neq \boldsymbol{e}_j$ unless $f(x_{ji}) = 1$.

Theorem F.2.3 If A is nonsingular but has generalised inverse A^-, then

$$AA^-\boldsymbol{v} \;=\; \boldsymbol{v}.$$

Proof For singular A, a generalised inverse of A is such that

$$\begin{aligned} AA^-A \;=\; A \;&=\; [\ \boldsymbol{a}_1 \quad \boldsymbol{a}_2 \quad \ldots \quad \boldsymbol{v} \quad \ldots \quad \boldsymbol{a}_p\ ] \\ &=\; [\ AA^-\boldsymbol{a}_1 \quad AA^-\boldsymbol{a}_2 \quad \ldots \quad AA^-\boldsymbol{v} \quad \ldots \quad AA^-\boldsymbol{a}_p\ ]. \end{aligned}$$

Therefore, if the j-th column is $\boldsymbol{v}$ then $AA^-\boldsymbol{v} = \boldsymbol{v}$. □

Corollary F.2.2 The scalar $\boldsymbol{v}^T A^-\boldsymbol{v}$ is the (j,j)-th element of A since

$$A \;=\; AA^-A \;=\; \begin{bmatrix} \boldsymbol{a}_1^T \\ \boldsymbol{a}_2^T \\ \vdots \\ \boldsymbol{v}^T \\ \vdots \\ \boldsymbol{a}_p^T \end{bmatrix} [\ A^-\boldsymbol{a}_1 \quad A^-\boldsymbol{a}_2 \quad \ldots \quad A^-\boldsymbol{v} \quad \ldots \quad A^-\boldsymbol{a}_p\ ].$$

Thus if $\boldsymbol{v} = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)$ and $A = \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)$ with

$$\boldsymbol{f}(\boldsymbol{x}_i) \;=\; \sum w_i [f(x_{1i}), f(x_{2i}), \ldots, 1, \ldots, f(x_{pi})]^T,$$

then

$$\boldsymbol{v}^T A^-\boldsymbol{v} \;=\; \sum w_i,$$

since $\sum w_i$ is the (j,j)-th element of A.

F.3 Equivalence of Designs for Linear Regression Models

F.3.1 Equivalence of D-optimum and D_s-optimum designs

Theorem F.3.1 For linear regression models with full parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, D-optimum and D_s-optimum designs are equivalent.

Proof The information matrix M for a linear regression model, and consequently its inverse M^{-1}, are block diagonal and can be written

$$M = \begin{bmatrix} M_{11} & \mathbf{0} \\ \mathbf{0} & M_{22} \end{bmatrix}, \qquad M^{-1} = \begin{bmatrix} M_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & M_{22}^{-1} \end{bmatrix}.$$

Since $M_{12} = M_{21}^T = \mathbf{0}$ in the partitioned information matrix of $\boldsymbol{\theta}$ (c.f. Section 5.1), the criterion function for D-optimality is given by

$$\begin{aligned} -\ln|M^{-1}| &= -\ln|M_{11}^{-1}| - \ln|M_{22}^{-1}| \\ &= \ln|(1/\sigma^2)\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)| + \ln(1/2\sigma^4). \end{aligned}$$

The derivative of the criterion function with respect to the design weights is

$$-\frac{d}{d\boldsymbol{w}} \ln|M^{-1}| = -\frac{d}{d\boldsymbol{w}} \ln|M_{11}^{-1}|,$$

since $M_{22} = 1/2\sigma^4$ is independent of the design weights $\boldsymbol{w}$. Hence, for linear regression models, the criterion function for D-optimality is maximised when the criterion function for D_s-optimality is maximised. Consequently, experiments can only be optimal for estimation of $\boldsymbol{\beta}$ and not σ^2. □

F.3.2 Equivalence of A-optimum and C-optimum designs

Theorem F.3.2 For linear regression models with full parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, A-optimum and C-optimum designs with $C^T = [I_s,\ \mathbf{0}]$ are equivalent.

Proof The proof follows in a similar vein to the above proof for D-optimality. The criterion function for A-optimality is given by

$$\begin{aligned} -\text{tr}\,\{M^{-1}\} &= -\text{tr}\,\{M_{11}^{-1}\} - \text{tr}\,\{M_{22}^{-1}\} \\ &= -\text{tr}\,\{(1/\sigma^2)\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\} - 2\sigma^4, \end{aligned}$$

with derivative

$$-\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M^{-1}\} = -\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M_{11}^{-1}\}.$$

□

Note that when considering the parameter vector $\boldsymbol{\theta} = \boldsymbol{\beta}$, a D_s-optimum design is used if interest is in good estimation of a subset of s of the $\boldsymbol{\beta}$ parameters. Here we are considering the extended parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. A D_s-optimum design in this case is used if interest is in good estimation of a subset of s of the $\boldsymbol{\theta}$ parameters.

F.4 Further Proofs for Stochastic Frontier Models

The following results pertain only to information matrices with nonsingular submatrix $F_\tau(\mu_a)$, which is the $(2,2)$ block of the information matrix associated with the $\boldsymbol{\tau}$ parameters. Nonsingularity of this $(2,2)$ block may occur if Methods 2 or 3 of Chapter 3 are used to approximate the information matrix. The $(2,2)$ block is singular under approximation Method 1.

F.4.1 Determinant criterion function

Theorem F.4.1 If the $(2,2)$ block ,$F_\tau(\mu_a)$, of approximated information matrix (6.2) is nonsingular, then D_A-optimum designs for the linear model

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + E, \qquad \mathbb{E}[E] = \mathbb{E}[-U], \tag{F.1}$$

are also D_A-optimum for the equivalent linear regression model with an intercept, given by

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + E, \qquad \mathbb{E}[E] = 0. \tag{F.2}$$

Proof The proof requires showing that

$$\frac{d}{d\boldsymbol{w}} \ln |M^{11}| = \frac{d}{d\boldsymbol{w}} \ln |M_{11}^{-1}|,$$

that is, that a design that is D_A-optimum for stochastic frontier model (4.9) is D_A-optimum, for an equivalent linear regression model. The criterion of D_A-optimality here has matrix $A = [I_p,\ \mathbf{0}]^T$, where $p = m + 1$, so that interest is in optimal estimation of the $\boldsymbol{\beta}$ parameters.

Part (i) of the proof:

The first part of the proof equates the M_{11} partitions of linear regression models and stochastic frontier models.

Let the matrix M, and its generalised inverse M^-, be partitioned as

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{12} & M_{22} \end{bmatrix}, \qquad M^- = \begin{bmatrix} M^{11} & M^{12} \\ M^{12} & M^{22} \end{bmatrix}.$$

For the log-linear stochastic production frontier model (4.9), let its approximated information matrix (6.2) be partitioned with

$$M_{11} = f_\beta(\mu_a) \sum w_i \boldsymbol{f}(\boldsymbol{x}_i) \boldsymbol{f}^T(\boldsymbol{x}_i),$$

$$\begin{aligned} M_{12} &= M_{21}^T, \\ M_{21} &= \boldsymbol{f}_{\beta,\tau}(\mu_a) \sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i), \\ M_{22} &= F_\tau(\mu_a). \end{aligned}$$

It follows that

$$\begin{aligned} -\ln|M_{11}^{-1}| &= p \ln f_\beta(\mu_a) + \ln\left|\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right|, \\ -\frac{d}{d\boldsymbol{w}} \ln|M_{11}^{-1}| &= \frac{d}{d\boldsymbol{w}} \ln\left|\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right|. \end{aligned}$$

From Section 5.1, the M_{11} block of the information matrix for a linear regression model with parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ is given by

$$M_{11} = \frac{1}{\sigma^2} \sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i).$$

Hence the D_A-optimum criterion function for optimal estimation of $\boldsymbol{\beta}$, and its derivative with respect to the weights, are

$$\begin{aligned} -\ln|M_{11}^{-1}| &= -p \ln \sigma^2 + \ln\left|\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right|, \\ -\frac{d}{d\boldsymbol{w}} \ln|M_{11}^{-1}| &= \frac{d}{d\boldsymbol{w}} \ln\left|\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right|. \end{aligned}$$

Since the M_{11} partition of the information matrix for a stochastic frontier model is proportional to the M_{11} partition of the information matrix for a linear regression model, the derivative of $\ln|M_{11}^{-1}|$ with respect to the weights are equal for both models.

Another equivalent argument is that the M_{11} element of the information matrix for the stochastic frontier model can be derived through a linear transformation of the design space for the linear regression model using transformation $\boldsymbol{g}(\boldsymbol{x}_i) = \sqrt{f_\beta(\mu_a)}\sigma I_p \boldsymbol{f}(\boldsymbol{x}_i)$. Therefore, by Theorem 6.1.1, $-\ln|M_{11}^{-1}|$ for the stochastic frontier model is maximised when $-\ln|M_{11}^{-1}|$ for the linear regression model is maximised.

Part (ii) of the proof:

The second part of the proof demonstrates that $|M^{11}|$ for the stochastic frontier model is proportional to $|M_{11}^{-1}|$.

From the results on inverses of partitioned information matrices given in Appendix E.1, if M_{22} and $(M_{11} - M_{12}M_{22}^{-1}M_{21})$ are nonsingular then

$$\begin{aligned} M^{11} &= \left[M_{11} - M_{12}M_{22}^{-1}M_{21}\right]^{-1} \\ &= \left[M_{11} - d\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i)\right]^{-1}, \end{aligned}$$

where $d = \boldsymbol{f}_{\beta,\tau}^T(\mu_a) \cdot [F_\tau(\mu_a)]^{-1} \cdot \boldsymbol{f}_{\beta,\tau}(\mu_a)$. Using equation (E.2) in Appendix E.2

$$M^{11} = M_{11}^{-1} + \frac{d\{M_{11}^{-1}\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\}\{M_{11}^{-1}\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\}^T}{1 - d\sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i)\{M_{11}^{-1}\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\}},$$

where $\{M_{11}^{-1}\}^T = M_{11}^{-1}$ since M_{11} is symmetric. By Theorem F.2.1, if the j-th element of $\boldsymbol{f}(\boldsymbol{x}_i)$ is 1, that is, if there is an intercept term in the model, then $f_\beta(\mu_a)\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)$ is the j-th column of $M_{11} = f_\beta(\mu_a)\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)$. Theorem F.2.2 then gives the result that $M_{11}^{-1}\sum w_i \boldsymbol{f}(\boldsymbol{x}_i) = f_\beta(\mu_a)^{-1}\boldsymbol{e}_j$. Substituting this into the equation above gives

$$\begin{aligned} M^{11} &= M_{11}^{-1} + \frac{df_\beta(\mu_a)^{-2}}{1 - df_\beta(\mu_a)^{-1}\sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{e}_j}\boldsymbol{e}_j\boldsymbol{e}_j^T \\ &= M_{11}^{-1} + \kappa f_\beta(\mu_a)^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^T, \end{aligned} \tag{F.3}$$

where $\sum w_i \boldsymbol{f}^T(\boldsymbol{x}_i)\boldsymbol{e}_j = 1$ by Corollary F.2.1, and $\kappa = \dfrac{df_\beta(\mu_a)^{-1}}{1 - df_\beta(\mu_a)^{-1}}$. The determinant can be derived using equation (E.1) in Appendix E.2 and is given by

$$\begin{aligned} |M^{11}| &= \left\{1 + \kappa f_\beta(\mu_a)^{-1}\boldsymbol{e}_j^T M_{11}\boldsymbol{e}_j\right\}|M_{11}^{-1}| \\ &= \{1 + \kappa\}\,|M_{11}^{-1}|, \end{aligned}$$

where $\boldsymbol{e}_j^T M_{11}\boldsymbol{e}_j = f_\beta(\mu_a)$ since the j-th element of $\boldsymbol{f}(\boldsymbol{x}_i)$ is 1. Taking negative logarithms on both sides gives the criterion function for D_A-optimality as

$$-\ln|M^{11}| = -\ln\{1 + \kappa\} - \ln|M_{11}^{-1}|.$$

The derivative with respect to the weights is then given by

$$-\frac{d}{d\boldsymbol{w}} \ln |M^{11}| \;=\; -\frac{d}{d\boldsymbol{w}} \ln |M_{11}^{-1}|,$$

since κ is independent of the design weights $\boldsymbol{w}$. Therefore a design that is D_A-optimum, for a linear regression model with nonzero intercept is also D_A-optimum for the equivalent stochastic frontier model, since, from part (i) of the proof, $-\frac{d}{d\boldsymbol{w}} \ln |M_{11}^{-1}|$ is equivalent for the two models. □

Remark Corollary 6.4.1, which is less technically complicated, would also be sufficient in proving the above theorem.

F.4.2 Trace criterion function

Theorem F.4.2 If the $(2,2)$ block ,$F_\tau(\mu_a)$, of approximated information matrix (6.2) is nonsingular, then C-optimum designs for the linear model (F.1) are also C-optimum for the equivalent linear regression model (F.2) with an intercept.

Proof The proof follows the proof above for the determinant criterion but requires showing that

$$\frac{d}{d\boldsymbol{w}}\text{tr } M^{11} \;=\; \frac{d}{d\boldsymbol{w}}\text{tr } M_{11}^{-1}.$$

The criterion of C-optimality here has matrix $A = [I_p,\ \mathbf{0}]^T$, where $p = m+1$, so that interest is in optimal estimation of the $\boldsymbol{\beta}$ parameters.

Part (i) of the proof:

For the log-linear stochastic production frontier model (4.9)

$$\begin{aligned} -\text{tr } \{M_{11}^{-1}\} &= -f_\beta(\mu_a)^{-1}\text{tr } \left\{ \left[\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) \right]^{-1} \right\}, \\ -\frac{d}{d\boldsymbol{w}}\text{tr } \{M_{11}^{-1}\} &= -f_\beta(\mu_a)^{-1}\frac{d}{d\boldsymbol{w}}\text{tr } \left\{ \left[\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i) \right]^{-1} \right\}. \end{aligned}$$

The C-optimum criterion function for optimal estimation of $\boldsymbol{\beta}$ in a linear regression model, and its derivative with respect to the weights, are

$$\begin{aligned} -\text{tr}\,\{M_{11}^{-1}\} &= -\sigma^2\text{tr}\,\left\{\left[\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right]^{-1}\right\}, \\ -\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M_{11}^{-1}\} &= -\sigma^2\frac{d}{d\boldsymbol{w}}\text{tr}\,\left\{\left[\sum w_i \boldsymbol{f}(\boldsymbol{x}_i)\boldsymbol{f}^T(\boldsymbol{x}_i)\right]^{-1}\right\}. \end{aligned}$$

Hence the derivative of $-\text{tr}\,\{M_{11}^{-1}\}$ with respect to the weights for both models are maximised by the same design.

Part (ii) of the proof:

From equation (F.3)

$$\begin{aligned} -\text{tr}\,\{M^{11}\} &= -\text{tr}\,\{M_{11}^{-1} + \kappa f_\beta(\mu_a)^{-1}\boldsymbol{e}_j\boldsymbol{e}_j^T\} \\ &= -\text{tr}\,\{M_{11}^{-1}\} + \kappa f_\beta(\mu_a)^{-1}. \end{aligned}$$

The derivative with respect to the weights is then given by

$$-\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M^{11}\} = -\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M_{11}^{-1}\},$$

since κ is independent of the design weights $\boldsymbol{w}$. Therefore a design that is C-optimum, for a linear regression model with nonzero intercept is also C-optimum for the equivalent stochastic frontier model, since, from part (i) of the proof, $-\frac{d}{d\boldsymbol{w}}\text{tr}\,\{M_{11}^{-1}\}$ is equivalent for the two models. □

Remark The pivotal assumption in Theorems F.4.1 and F.4.2 is nonsingularity of the $(2,2)$ block of the information matrix. The results on the trace criterion in Section 6.5 do not agree with Theorem F.4.2 since the $(2,2)$ block of the information matrix is singular for the non-approximated information matrix. It is also singular when the information matrix is approximated using Method 1 of Chapter 3. Use of approximation methods 2 and 3 in Chapter 3, for which the theorem above holds, is not advisable since positive definiteness of the information

matrix cannot be guaranteed. Additionally, it is not desirable for the optimum design using an approximated information matrix to differ greatly to the optimum design using the non-approximated information matrix.